Feature Selection for High Dimensional Causal Inference

Rui Lu

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Feature Selection for High Dimensional Causal Inference

Rui Lu

Selecting an appropriate set for confounding control is essential for causal inference. The strong ignorability is a strong assumption. With observational data, researchers are unsure the strong ignorability assumption holds. To reduce the possibility of the bias caused by unmeasured confounders, one solution is to include the widest range of pre-treatment covariates, which has been demonstrated to be problematic. Subjective knowledge-based covariate screening is a common approach that has been applied widely. However, under high dimensional settings, it becomes difficult for domain experts to screen thousands of covariates. Machine learning based automatic causal estimation makes it possible for high dimensional causal estimation. While the theoretical properties of these techniques are desirable, they are only necessarily applicable asymptotically (i.e., requiring large sample sizes to be guaranteed to hold), and their performance in smaller samples is sometimes less clear. Data-based pre-processing approaches may fill this gap. Nevertheless, there is no clear guidance on when and how covariate selection should be involved in high dimensional causal estimation.

In this dissertation, I address the above issues by (a) providing a classification scheme for major causal covariate selections methods (b) extending causal covariate selection framework (c) conducting a comprehensive empirical Monte Carlo simulation study to illustrate theoretical properties of causal covariate selection and estimation methods, and (d) following-up with a case study to compare different covariate selection approaches in a real data testing ground.

Under small sample and/or high dimensional settings, study results indicate choosing an appropriate covariate selection method as pre-processing tool is necessary for causal estimation. Under relatively large sample and low dimensional settings, covariate selection is not necessary for machine learning based automatic causal estimation. Careful pre-processing guided by subjective knowledge is essential.

Table of Contents

List of Charts

List of Tables

Acknowledgments

This dissertation was completed during the COVID-19 pandemic. It had been decades since such a difficult period occurred in the United States. During this challenging time, there were numerous individuals that added light to my life while studying at Columbia University. I am thankful for my grandparents, Jiefu Lu and Zehui Sun, who passed away at the time I was writing my dissertation. I will always remember grandpa JieFu's wish for me to contribute to society as an active scholar. I will always remember grandma Zehui's recipes for cooking delicious food, which has been essential the during pandemic. I am thankful for my cousin Zhiwei Wang, who passed away during the second year of my Ph.D. Inspired by Zhiwei, I am determined to be a causal statistician.

I am thankful for my parents, Ping Wang and Chong Lu who always support me and encourage me during difficult times. I am also grateful for my academic parents Dr. Bryan Keller and Dr. Jennifer Hill. My advisor Dr. Keller taught me, step by step, how to construct my first code and how to write academic papers during my Ph.D. The training I received from you has been one of the most valuable experiences of my life. My mentor Dr. Jennifer Hill opened the magic door for me to enter the world of causal inference during my work at NYU. Thank you for your generosity in sharing your thoughts and your careful guidance regarding my works and researches. I also want to thank my committee members: Dr. James Corter, Dr. Caleb Miles and Dr. Sarah Cohodes for their helpful comments and support on my dissertation.

In memory of Jiefu and Zehui

Chapter 1: Introduction

Statisticians often refer to randomized experiments as the "gold standard" for determining the causal effect of a binary intervention on an outcome. Randomized experiments are designed to eliminate the systematic differences of measured and unmeasured covariates across treatment groups. However, in many educational settings, fully randomized assignments may not always be ethically or practically possible. For example, it is unethical to randomly assign students into different schools or for participants to receive curriculum of unequal quality. In such cases, observational studies may be used to estimate the causal effect. Instead of being randomly assigned, the subjects in observational studies would be self-selected or otherwise selected into treatment groups. The implications of groups receiving different treatments is that the subjects in each group are likely to be different in ways that are relevant to the outcome. The variables that relate both to the outcome and treatment are called confounding variables. Therefore, if there are differences in outcomes across treatment groups, researchers would not be able to determine if the outcomes are caused by the treatment or the confounding variables. The bias caused by self-selection is often referred to as selection bias.

One strategy for estimating average causal effects with observational data is through statistical conditioning. Conditioning methods require either a modeling of response surfaces (i.e., regression modeling), a propensity score model or both (i.e., doubly robust methods). To control the selection bias, either the treatment or control model would need to be correctly specified while also including, at minimum, all the confounders. To reduce the bias caused by unmeasured confounders, one solution is to include the widest range of pre-treatment covariates (called the "kitchen sink" approach). Through the development of survey techniques and widely available educational

databases. For instance, it is common for an educational data set to include hundreds or even thousands of covariates. Crude controlling for all pretreatment covariates may result in the following problems: (1) inability to fit the model due to the curse of dimensionality (Schnitzer, Lok & Gruber, 2016); (2) insufficient overlap across treatment groups which may cause estimation to be prone to bias (Shafer & Kang, 2008); (3) estimator's variance inflation caused by including strong predictors of the treatment or many covariates unrelated to outcome and treatment (Greenland, 2008; Patrick et al., 2011; Schisterman et al., 2009); and (4) inclusion of covariates that would introduce bias (i.e., collider) or amplify it (i.e., instrument variable) (Myers et al, 2011; Pearl, 2010, 2011). Researchers face the condition that either innocently omitted or indiscriminately included pretreatment covariates may cause problems for causal estimation. Therefore, some type of feature selection is needed to achieve unbiased, efficient estimation.

This dissertation aims to identify and resolve some of the existing problems for causal feature selection. Although there are other methods outside of conditioning strategies that permit the identification and estimation of average causal effects, I limit the scope of my dissertation to conditioning strategies because of the special role that covariates play in conditioning. In the remainder of this chapter, I will review two fundamental frameworks that allow for the identification of causal effects: the potential outcomes notation and causal graphs. Afterward, causal feature selection will be defined. Unresolved problems in causal feature selection will be framed into three research questions. I will conclude the chapter with an overview of this dissertation.

1.1 The Potential Outcomes Notation

Potential outcomes notation is widely used in the statistical community to represent causation. Neyman (1923) introduced this concept in his agricultural field experiment. Later, Rubin (1974,

1977, 1978, 1980) extended it to observational studies (Sobel, 2005). To establish causality, two elements are important: (1) each experimental unit must be *potentially exposed* to each of the treatment conditions, and (2) there must be a *temporal division* between cause and effect (Holland, 1986).

The causal effect for a given individual may be defined as the comparison between the effects of the factual (what has happened to the individual) and the counterfactual (what would have happened to the individual). Let T_i denote the treatment assignment for individual i, with $T_i = 1$ indicating assignment to treatment (or identically, exposure) and $T_i = 0$ indicating assignment to control (or another treatment). Each individual subject has two potential outcomes, denoted by Y_i^1 and Y_i^0 respectively. The individual treatment effect (ITE) for individual i is defined as the difference between two hypothetically obtained outcomes:

$$ITE_i = Y_i^1 - Y_i^0. \tag{1.1}$$

Under this notation, there are two classes of variables: the pretreatment variable X, which is measured before the treatment has been assigned, and the post treatment variable Y, which is measured after exposure to the treatment.

1.1.1 Average Treatment Effect (ATE)

We cannot observe the value of the response variable under both treatments, but only through the observed outcome, $Y_i = Y_i^1(T_i) + Y_i^0(1 - T_i)$. This is known as the fundamental problem of causal inference (Holland, 1986; Rubin, 1976). Therefore, the individual treatment effect is never observable. For this reason, in many cases, the quantity of interest is the average treatment effect (ATE):

$$\tau_{ATE} = E(Y_i^1 - Y_i^0) = E(Y_i^1) - E(Y_i^0) \tag{1.2}$$

or the average treatment effect on the treated (ATT):

$$\tau_{ATT} = E(Y_i^1 - Y_i^0 | T_i = 1). \quad (1.3)$$

1.1.2 Assumptions for Identifying the ATE in the Observational Studies

In this example, X_i, Y_i^0, Y_i^1 are vectors of individual values. By using potential outcome notation, the assignment mechanism may be written as:

$$P\left(T_i \middle| X_i, Y_i^{\,0}, Y_i^1\right). \quad (1.4)$$

The only random variable in 1.4 is T; all other variables are fixed (Rubin, 2005). A critical property is called *ignorability:*

$$\{Y_i^1, Y_i^0\} \perp T \,|X. \quad (1.5)$$

Under assumption 1.5, it is justifiable to ignore the missing values (unobserved potential outcomes). Therefore, the assignment mechanism could be rewritten as: $P(T_i \middle| X_i, Y_i^0, Y_i^1) = P(T_i|X_i)$. Another important property is that the unit level probabilities of the treatment is between 0 and 1:

$$0 < P(T_i = 1|X_i) < 1. \quad (1.6)$$

Intuitively, every unit has a positive chance of being assigned into either treatment or control conditions (this is referred to as *positivity*). Combining assumptions 1.5 and 1.6, it becomes the stronger version of *ignorability*, called *strong ignorability*. Hence, $P(T_i \middle| X_i, Y_i^0, Y_i^1) = P(T_i|X_i)$. Rosenbaum and Rubin (1983) named $e(X_i) = P(T_i = 1|X_i))$ the *propensity score*, which collapses covariates into a scalar, making it easier to condition the propensity score rather than all covariates. Rosenbaum and Rubin (1983) proved that if *ignorability* holds, then independence between the potential outcomes and the treatment assignment may be attained by conditioning on the unidimensional propensity score. That is,

$$\{Y_i^1, Y_i^0\} \perp T_i \,|X_i \Rightarrow \{Y_i^1, Y_i^0\} \perp T_i \,|e(X_i). \quad (1.7)$$

In assumptions 1.2 and 1.3, the expectations are taken over the joint distribution of P (T, Y^1, Y^0). From this perspective, each unit is assumed to receive the same version of treatment and the value of each potential outcome is independent of their particular assignment pattern in *T*. This is called the *stable unit treatment value assumption* (SUTVA; Rubin, 1978, 1980). Different from a randomized experiment, in observational studies, the potential outcomes are typically not independent of the treatment assignment, such that $E(Y_i^1) \neq E(Y_i|T_i = 1)$ and $E(Y_i^0) \neq E(Y_i|T_i = 0)$. Direct estimation of the ATE by taking group sample averages may lead to selection bias. By satisfying the *strong ignorability assumption* and SUTVA, the average causal effect can be unbiasedly estimated though conditioning strategies. That is modeling one or both of $P(Y_i|X_i, T_i)$ and $P(T_i|X_i)$.

1.2 The Graphic Approach

In causal inference, a directed acyclic graph (DAG) is a graphic representation of how variables are causally connected. DAGs were initially proposed by Pearl (1988, 1995, 2009) and Spirtes et al. (1993) and Spirtes (2001). In this framework, DAGs are given two distinct functions. In the first, a DAG is used to represent the skeleton of causal structures. In the second, a DAG is used to represent underlying probability distributions.

1.2.1 Skeleton of Causal Structures

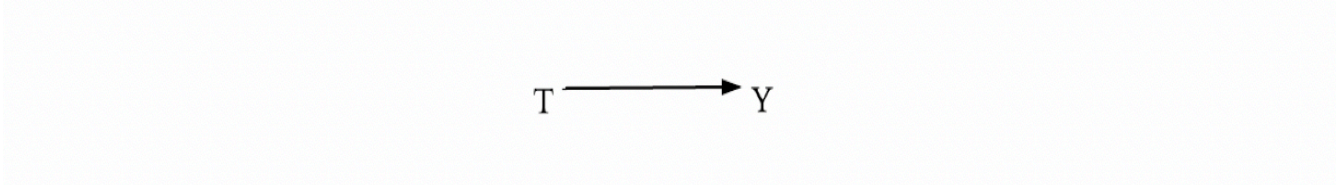


Figure 1.1: An example of a DAG.

DAGs that are interpreted causally are called causal graphs. There are three construction components for DAGs: nodes (vertices), directed arrows, and missing arrows. Nodes represent

random variables in the causal model. Arrows represent the possibility of a direct causal effect between variables and their order in time. For example, in Figure 1.1, the directed arrow between T and Y means the following: (1) treatment (T) occurred before response (Y), and (2) T exerts causal effect on Y. A missing arrow indicates no direct causal effect between two variables in a defined population.

A path is composed by connecting directed edges. For instance, in Figure1.1, T and Y are connected by a path (T→ Y). In a causal DAG, a variable directly caused by a given variable is referred to as the child. The direct cause of the variable is called its parent. All variables that are directly or indirectly caused by a given variable are referred to as a descendant. No variable could be its own descendant in the graph. In other words, there should be no cycles (represented in Figure 1.2).

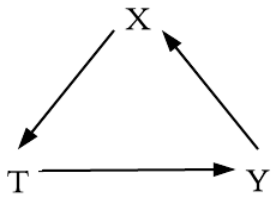


Figure 1.2: Example of a directed graph that includes a cycle.

1.2.2 Probability Distribution and Conditional Independence

The second function of a DAG is to be used to represent the joint distribution of its nodes if the probability distribution could be Markov factorized as:

$$P(v) = \prod_{i=1:n} P(v_i|pa_i), \tag{1.8}$$

where v denotes the vector of all notes in DAGs and pa is their parents (Pearl, 1995). In other words, each variable is independent of all variables in the past and its non-descendants given its parents (also called the *Causal Markov Assumption*).

When three variables (X, Y, T) are connected by a path in a DAG, there are three basic relationships. The first is called *mediation* (T→ X→ Y) (left graph of Figure 1.3) where the relationship between T and Y is mediated by X. The second one is called *mutual dependence* or c*onfounding* (T← X→ Y) (middle graph of Figure 1.3) in which X (confounder) is the parent for both T and Y. The third one is called *mutual causation* or *colliding* (T→ X ← Y), where X (*collider*) is the common effect for both T and Y (right graph of Figure 1.3). A path between two variables T and Y is said to be *d-separated* (blocked or closed) by another set of variables T if:
(1) conditioned on the path that contains a non-collider or (2) not conditioned on the path that contains a collider (Pearl, 1995; Pearl, 2009).

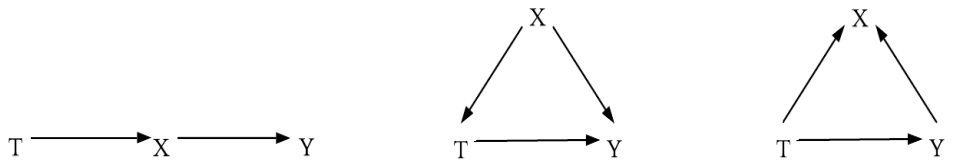


Figure 1.3: Three possible relationships for three variables in a DAG.

By assuming there are no conditional independence relations other than those entailed by the *Causal Markov Assumption*, it could be deduced that *d-separation* implies statistical independence (Scheines, 1997). For example, in the middle graph of Figure 1.3, T and Y are d-separated after a condition on X. Under the *Causal Markov assumption*, it is equivalent to conditional independence: $T \perp Y|X$.

1.2.3 Identification

A causal DAG allows for predicting the causal effect of an intervention (denoted by do(T=t)). Given two disjointed sets of variables, T and Y, the causal effect of T on Y is denoted as:

$$P(Y|do(T=t)) = \begin{cases} \frac{P(T,Y)}{P(T=t|pa_t)}, & T=t \\ 0, & T \neq t \end{cases}. \quad (1.9)$$

The ATE could be denoted as E(Y|do(T=1))-E(Y|do(T=0)) under the assumption that the individual level of causal effect is defined by the individual level difference induced by the hypothetical intervention. One of the major strengths of casual DAGs is that they may be used through a series of rules (called the *backdoor path criteria*) to determine the sufficient covariate set that satisfies the ignorability assumption. For instance, in the middle graph of Figure 1.3, conditioning on covariate set X could block all the backdoor paths from T to Y. This means no nodes in T are descendants of X, and X blocks every path between T and Y that contains an arrow into Y. If X satisfies the backdoor path criteria, the interventional distribution is given by:

$$\begin{aligned} P(Y|do(T=t)) &= \sum_x P(Y|T, X=x)P(X=x) \\ &= \sum_x \frac{P(X=x,T,Y)}{P(T|X=x)} , \end{aligned} \tag{1.10}$$

where $P(T|X=x)$ is the propensity score. The formula (1.10) is the "inverse probability weighting formula" for causal effects estimation. Therefore, the back-door path criterion may be viewed as a generalized form of the *ignorable assumption* (Pearl, 1995).

1.3 Causal Feature Selection

In this dissertation, I refer to the term *feature* as covariates or a weighted/transformed version of covariates. Causal feature selection could be defined as the process of identifying relevant features and discarding irrelevant ones with the aim of obtaining a subset of features for the purposes of controlling for confounding bias and enhancing estimation efficiency. Causal feature selection is composed of explicit causal covariate selection and implicit causal covariate weighting. Covariate selection identifies specific covariates for control whereas covariate weighting assigns weights to different covariates. Covariate selection could be also viewed as a special case of causal covariate weighting with selection weights of 0 and 1.

1.3.1.1 Causal Covariate Selection

A central aim for causal covariate selection is to identify a set of covariates that is sufficient for confounding adjustment. A covariate set satisfying the *ignorability* assumption would provide a sufficient adjustment set. This is different from covariate selection for prediction, where accuracy is the main concern. The same is true for descriptive modeling, where sparse representation of association structure is the focus. In the following paragraph I will define causal covariate selection.

First, I assume that *ignorability* holds either (a) with the set of all measured pretreatment covariates, or (b) does not hold with the complete set of pretreatment covariates, however, there is a subset for which it could hold. Scenario (b) is possible with, for example, with the presence of collider variables. In this case, I define covariate feature selection as a procedure that aims to identify a subset $X_s \subseteq X$ for which *ignorability* holds, efficiency of estimation (i.e., $\mathrm{Var}(\tau)$) is increased, and positivity is satisfied. That is, $\{Y^1, Y^0\} \perp T \mid X \Rightarrow \{Y^1, Y^0\} \perp T \mid X_s$ such that $\mathrm{Var}(\tau_{X_s}) < \mathrm{Var}(\tau_X)$ and $0 < p(T|X_s) < 1$. I am not considering the case when *ignorability* is not satisfied since valid causal inference could not be drawn. In this case, dimension reduction by feature selection still plays an important role for bounding problems or sensitivity analysis (see Kennedy & Balakrishnan, 2018).

1.3.2 Covariate Weighting by Regularization

To begin, I will distinguish the terms covariate selection and regularization. Covariate selection is the procedure in which a specific covariate or a subset of covariates are identified. In contrast, to deal with high dimensional observations, regularization is a group of estimation methods that modify objective function by introducing penalties into estimation procedure (Bickel et al., 2006). The typical form of these methods is:

$$\hat{\beta} = \mathrm{argmin}_{\beta} \{l(\beta) + \lambda \sum_{j=1}^{p} p_{\lambda}(\beta_j)\}, \tag{1.11}$$

where β is the vector of the regression coefficients; p_λ is the penalty function and l is the loss function. From the Bayesian perspective, regularization (1.11) is achieved by imposing prior model parameters. The typical penalization functions are Lasso (Tibshirani, 1996), elastic net (Zou & Hastie, 2005) and SCAD (Fan & Li, 2001).

Under satisfaction of the *strong ignorability* assumption, regularization could shrink some of the coefficients of the covariates to be zero or re-weight certain covariates. In causal inference, regularization is widely adapted for estimation of the propensity score (McCaffery, Rideway, & Morral, 2004; Ning, Peng & Imai, 2017); response surface (Hill, 2011) and doubly-robust estimation (Shortreed & Ertefaie, 2017). In certain cases, regularization methods could be used for covariate selection. Furthermore, covariate selection could be embedded in high dimensional regularization approaches (e.g., Collaborative-controlled Lasso by Ju et. al, 2019).

1.4 Statement of the Problems

1.4.1 Unclear Classification

There is a rapid increase in the amount of recent causal inference publications with the key words "covariate selection," "confounding control," "high dimensional estimation," and "model selection". They are all related to causal feature selection but come from different perspectives. There are papers focused on the types of covariates that should be included/excluded (Austin et al, 2007; Bhattacharya & Vogt, 2007; Brookhart et al, 2010; Middleton et al., 2016; Steiner et al., 2010); establishing selection theoretical frameworks or selection algorithms (De luna, Waernbaum, & Richardson, 2011; Maathuis & Colombo, 2013; VanderWeele & Shpitser, 2011); and methods to implement covariate selection and/or causal estimation (Hill, 2011; Persson et al., 2017; Wilson & Reich, 2014).

Researchers from different camps often have disparate philosophies on design and favor a variety of feature selection methods. For example, researchers who favor causal graphic approaches to causal identification may have different attitudes towards the use of causal graphs for feature selection when compared with researchers who prefer approaches to causal identification grounded in potential outcomes. Without a clear understanding of the primary approaches for causal feature selection, researchers will find it challenging to identify the best strategy to use based on research needs and to justify why they selected a specific approach.

1.4.2 Unfair Comparisons with Few Methods

Most publications focus on proposing new selection methods. Simulation tends to favor the proposed approach for a few reasons. First, authors have a better understanding of their own proposed methods compared to competitors. As a result, the best tuning parameters would be set for the proposed approach. For selected competitors, only the "vanilla" version would be used. Second, the evaluation metrics that authors select tend to favor the proposed methods. For instance, some papers focus more on bias rather than confidence interval coverage. There are only a few papers focused on comparing covariate selection/regularization methods. To the best of my knowledge, only one paper does not propose a new method (see Witte & Didelez, 2018) that focuses on comparing different causal covariate selection approaches through simulation. Nevertheless, this paper only primarily covers binary outcomes under simple simulation settings.

1.4.3 Difficult to Identify Appropriate Methods that Calibrate to Real Life

There are certain methods that are only proposed for theoretical reasons, which in turn, are difficult to implement in real-world settings. In addition, due to design characteristics, some methods are more suitable for a specific range of sample sizes or dimensions of the covariate set. For example, maximum likelihood-based methods, which often rely on asymptotic properties, may

not perform well with a small data set. Additionally, a cross-validation or permutation-based approach will require a relatively long computation time, which may not be suitable for a very large data set. Other factors, such as the functional forms of assignment model/response surfaces, correlation among covariates and signal to noise ratio also impact the successful implementation of a method. However, there is no clear suggestion mentioned in previous studies regarding how to best select appropriate methods based on the characteristics of the real data set.

1.5 Dissertation Objective

In this dissertation, I will address the above issues by (a) reviewing the literature to classify causal covariate selection methods, (b) providing a comprehensive empirical Monte Carlo simulation study to illustrate theoretical properties of causal covariate selection and estimation methods, and follow-up with a case study to compare different covariate selection approaches in a real data testing ground, (c) introducing new extensions of covariate selection methods to resolve problems that existed in previous methods.

In this study, I will explore the following questions:

Question 1. What is the classification scheme for major causal covariate selection methods?

1a. What are the primary approaches for covariate selection?

1b. What are the pros and cons of primary covariate selection methods?

1c. What are the existing problems?

Question 2. What are possible new extensions for covariate preprocessing that may improve upon the drawbacks of existing methods?

2a. What are these new extensions?

2b. Does the empirical and theoretical evidence base support these new extensions?

Question 3. What are the factors that influence performance of covariate selection approaches for ATE estimation?

3a. Are there any differences between feature selection approaches across different sample sizes?

3b. Are there differences with a higher dimensionality of covariate sets?

3c. Are there any differences for different covariate selection criteria?

3d. Is covariate selection still necessary when automatic estimation methods are used?

The remainder of this dissertation will be organized as follows: In Chapter 2, I will provide an overview and classification on causal feature selection approaches by conducting a literature review. Afterward, I will identify the primary feature selection methods for current research that would feasibly perform well in both simulation and real case studies. These methods also have the potential to be adapted to applied research. In Chapter 3, I will introduce recent developed extensions of methods motivated by exploring preliminary performances of existing methods via simulation and illustrate the design of simulation and empirical studies. In Chapter 4, I will provide the simulation and empirical study results. In Chapter 5, I will summarize findings in the simulation and empirical studies and discuss the implications regarding further research.

Chapter 2: Literature Review

This chapter consists of three sections. First, I will provide a broad overview of causal feature selection strategies. I will then present three primary approaches (i.e., filter approach, wrapper approach and regularization approach) alongside discussions of their advantages and limitations. I will also discuss methods that could be implemented in future research. In the final section, I will summarize existing problems in causal feature selection, including arguments from different perspectives. I will then turn these problems into simulation "knobs" for an empirical Monte Carlo simulation (EMCS) design and rationale for extensions to the methods that will be proposed in Chapter 3.

2.1 Overview of Causal Feature Selection Strategies

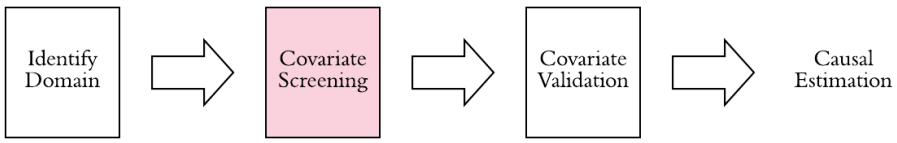


Figure 2.1: Flow chart for causal feature selection.

Sauer et al. (2013) suggested that a practical covariate selection should combine a prior covariate selection based on researchers' subjective knowledge about the causal relationship with empirical covariate screening. The procedure for covariate feature selection has been summarized in Figure 2.1. Subjective knowledge plays an important role for feature selection in causal inference. The first step always starts with domain experts who identify the covariates domains such as social economic status (SES) and educational background. Within each domain, there are several covariates. Empirical covariate selection methods discussed below could be used to

decrease the number of covariates within each domain. After screening, domain experts should validate the selected covariates based on subjective knowledge before moving towards causal estimation.

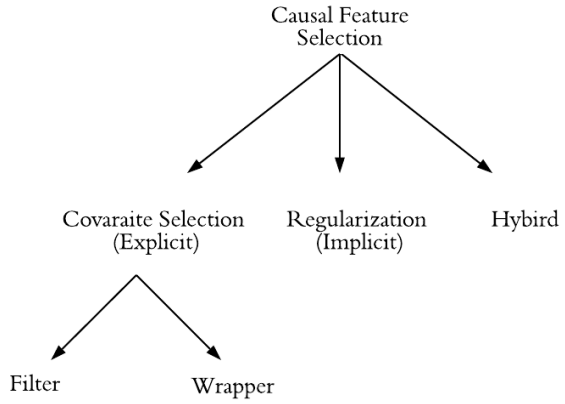


Figure 2.2: Classification of causal feature selection strategies.

There are three primary types of feature screening strategies (Figure 2.2) that are used frequently in causal inference: covariate selection, regularization (also called embedded methods), and a combination of the two (deemed the hybrid approach). Covariate selection is characterized by the explicit identification of a specific covariate set and is usually implemented as a pre-processing step before any causal response surface or propensity score modeling. In contrast, with regularization approaches, the feature selection and model estimation co-occur such that the training procedure arises implicitly by up-weighting significant features and down-weighting insignificant ones. The hybrid approach is any type of combination of covariate selection and regularization.

2.1.1 Classification of Causal Covariate Selection

Causal covariate selection methods could be further classified based on: (1) covariate search methods: different search mechanisms; (2) covariate selection criterion: choice of covariate set

for confounding control and (3) objective function (i.e., propensity score model, outcome model etc.).

2.1.1.1 Feature Search Mechanism

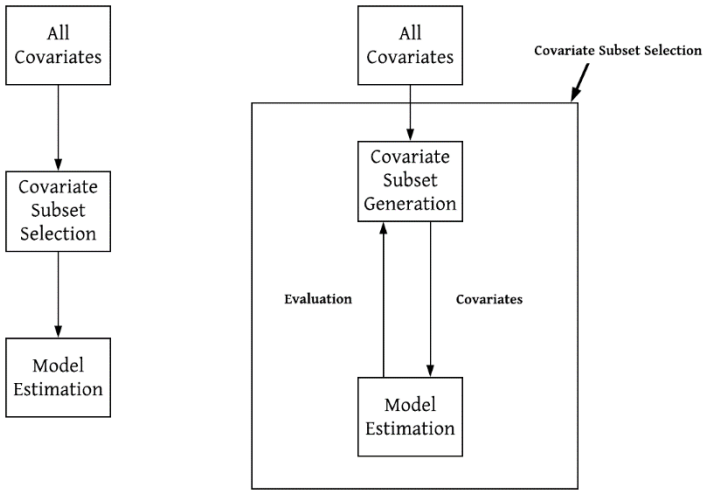


Figure 2.3: Filter method (left) and wrapper method (right).

Witte and Didelez (2018) distinguished between two types of general covariate selection mechanisms: pre-adjustment and wrapper (Figure 2.3). Pre-adjustment methods are also referred to as "filter" methods in the machine learning community (Kohavi & John, 1997).

Filter methods identify relevant covariates using variable ranking techniques (e.g., information criteria, p-values, heuristics, etc.). The selection is carried out as a two-step procedure. First, an appropriate ranking criterion is applied to score and rank the covariates. Afterwards, a threshold (e.g. p-values) is used to filter out relevant covariates. The selection procedure does not involve any kind of repeated estimation of models such as stepwise selection or cross-validation based selection.

One example of a filter method in causal inference is the Bayesian network approach proposed by Häggström (2018). It identifies the various types of subsets for confounding control by using

conditional independence testing with mutual information. One advantage of filter methods is that the selection process is not impacted by model estimation. Therefore, it provides a safeguard against researcher discretion and is less computationally expensive than wrapper methods (Witte & Didelez, 2018). Nevertheless, one of the concerns for filter methods is that they might include irrelevant covariates and exclude relevant covariates due to a Type I error of statistical tests (Khan & Quadri, 2013).

The term "wrapper" comes from the concept that the selection process is wrapped around the estimation procedure. For wrapper methods, the covariate selection cannot be separated from the estimation procedure. Note in Figure 2.3 the estimation process is carried out repeatedly by searching over spaces of covariate subsets. One set of covariates is selected in a stochastic way by optimizing pre-defined model evaluation criteria (often quality of fit such as MSE, AIC and BIC). The relevance of each covariate is determined by its importance for model fit. In causal inference, the change-in-estimate (CIE) method is a representative example of a wrapper method: a benchmark treatment effect is estimated by using all covariates. Afterwards, covariates are gradually removed until no further removal will result in change in estimate of more than pre-defined cutoff values in comparison to the bench mark estimate (Maldonado & Greenland, 1993; Mickey & Greenland, 1989).

Other examples of wrapper methods are stepwise regression, cross-validation and backward-elimination. Some researchers (i.e., Vansteelandt, Bekaert, & Claeskens, 2012) recommend wrapper methods because the stochastic selection procedure accounts for the uncertainty involved in covariate selection, which is often ignored by filter methods. However, the major drawbacks of wrapper methods are that they are often computationally intensive and run the risk of overfitting.

2.1.1.2 Feature Selection Criterion

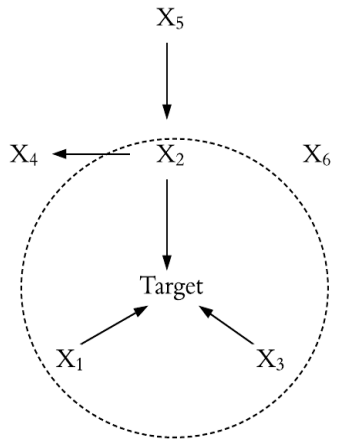


Figure 2.4: Covariate relevance; Dashed line indicate Markov blanket for target variable.

The first step is to define covariate relevance based on the target variable. John, Kohavi and Pfleger (1994) organized covariate relevance into three categories: strongly relevant, weakly relevant and irrelevant. Let S denote subsets such that exclude X_i from X. A feature X_i is:

- Strongly relevant if and only if $p(\text{Target}|\, X_i, S) \neq p(\text{Target}|S)$. For example, $\{X_1, X_2\ X_3\}$ in Figure 2.4. All strongly relevant covariates form a Markov blanket for the target variable.
- Weakly relevant if and only if $p(\text{Target}|\, X_i, S) = p(\text{Target}|\, S)$ and there exits $S' \subseteq S$ such that $p(\text{target}|\, X_i, S') \neq p(\text{Target}|\, S')$. For example, $\{X_4, X_5\}$ in Figure 2.4. A weakly relevant covariate is also called a proxy variable for its correlated covariates.
- Irrelevant if and only if for all $S' \subseteq S$ such that $p(\text{Target}|\, X_i, S') = p(\text{Target}|\, S')$. For example, $\{X_6\}$ in Figure 2.4.

The next step is to categorize the relationship between strongly relevant covariates for a target variable (i.e., treatment status (T) and outcome variable (Y)) in a causal DAG and understand their roles in causal estimation.

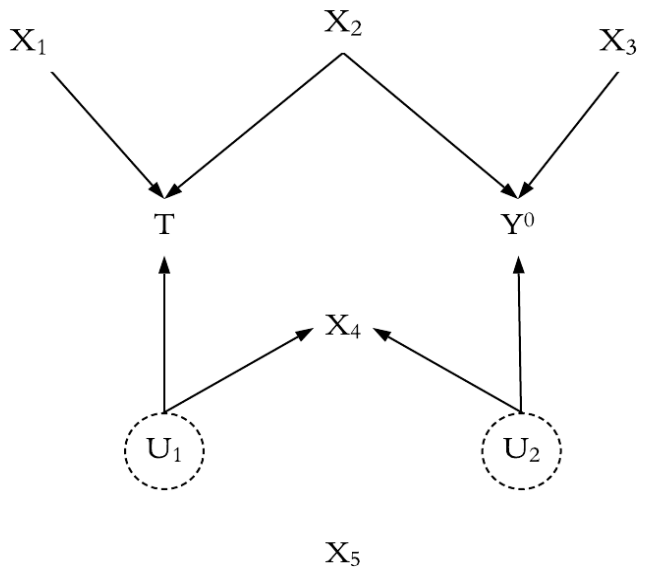


Figure 2.5: A causal DAG with different types of covariates.

The causal DAG in Figure 2.5 contains common types of covariates that are of interest in causal inference applications. Note that X_1, X_2, X_3, X_4 and X_5 are all observed covariates; U_1 and U_2 are unobserved.

Risk factors (X_3 in Figure 2.5) are predictors of the outcome. These have no influence on the treatment status. Controlling for such covariates will gain estimation efficiency without introducing estimation bias (Austin et al., 2007; Brookhart et al., 2010; Steiner et al., 2010; Shortreed & Ertefaie, 2017).

Confounders (X_2 in Figure 2.5) are predictors of both the exposure (T) and outcome (Y). Failing to control confounders would bias the non-experimental treatment estimation (Brookhart et. al 2010; Sauer et.al, 2013).

Instrumental variables are defined as a covariate that only influence the outcome through exposure (X_1 in Figure 2.5). If strong ignorability holds with the set of observed covariates, including (i.e., conditioning), an instrumental variable would not introduce bias but would inflate

the variance of the estimator (Brookhart et.al, 2006; Bhattacharya & Vogt, 2007). If strong ignorability does not hold, adjusting for an instrumental variable may amplify the impact of the exposure outcome confounding (Ding, VanderWeele & Robins, 2017; Middlton et al., 2016; Sauer et al., 2013).

Including *colliders* (X_4 in Figure 2.5) that are common effects of two independent covariates (U_1 and U_2) could result in bias. This is referenced as M-colliding bias (Greenland & Robins, 1986). Controlling for colliders can result in the violation of the *strong ignorability assumption*. Conditioning on X_4 would introduce confounding between T and Y by opening a path through the collider.

S*purious variables* (also called noise variables) are variables that have no direct causal effect on variables in the DAG (X_5 In Figure 2.5). Including such variables will not cause bias but can decrease efficiency (Persson et al., 2017; Shortreed & Ertefaie, 2017).

Based on causal graph theory, conditioning on covariate sets $\{X_2\}$, $\{X_1, X_2\}$, $\{X_2, X_3\}$, $\{X_1, X_2, X_3\}$ and $\{X_1, X_2, X_3, X_5\}$ would d-separate all paths between T and Y and therefore allow for identification of the average treatment effect. As a result, they are all sufficient sets for confounding control.

Based on different target sets, various selection criterions are defined. Control of all confounders ($\{X_2\}$) has been referred to as the *common cause criterion* (VanderWeele & Shipister, 2011). The approach of controlling for all predictors of treatment ($\{X_1, X_2\}$) has been called the *treatment criterion* (Witte & Didelez, 2018). Similarly, identifying all strong predictors of the outcome is called *outcome criterion* ($\{X_2, X_3\}$). Controlling for covariates that are either predictors of the treatment or outcome ($\{X_1, X_2, X_3\}$) is called the *disjunctive cause criterion* (VanderWeele & Shipitser, 2011).

To increase estimation efficiency, the optimal target set should only include confounders and risk factors $\{X_2, X_3\}$. Therefore, the outcome criterion is preferred and has begun to become widely adopted (Brookhart et. al 2010; Hill, 2011; Shortreed & Ertefaie, 2017). Nevertheless, there are others who believe that solely focusing on the outcome covariate relationship may exclude confounders that are weakly related to the outcome but strongly related to the treatment. Omitting such covariates would introduce bias (Vansteelandt, Bekaert, & Claeskens, 2012; Wilson & Reich, 2014). VanderWeele & Shipitser (2011) and VanderWeele (2019) suggest a two-step procedure, referred to as the *modified disjunctive cause criterion.* The first step identifies covariates that are predictors of either the outcome or the treatment ($\{X_1, X_2, X_3\}$). The second step further eliminates covariates that are not predictors of the outcome X_1 and include any proxy of the unmeasured confounders.

2.1.2 Classification of Causal Regularization

Regularization would help to prevent against overfitting of estimates of either the propensity function or the response surface with finite data. In this sense, regularization has the potential to improve ATE estimation. Regularization could be conducted by, for example, imposing a regularization term on regression coefficients, stopping early for learning algorithms, and using tree pruning or pre-defined Bayesian priors for parameters. For regularization, feature selection is embedded in the model training procedure. Most regularization methods select features by trading off the bias and variance for estimation. Regularization used in causal inference can be sub-divided based on different learning algorithms and the objective functions. I have classified three common types of regularization algorithms often used in causal inference.

1. **Penalized regression:** This approach focuses on using a penalized regression model for confounding control though direct regularization or Bayesian regularization through priors.

For example, post double selection (Belloni, Chernozhukov, & Hansen, 2014) and Bayesian regularized regression (Hahn et al., 2018).

2. **Regularized trees and ensembles:** This approach uses an adapted ensemble tree model for high dimensional causal estimation. Ensemble trees are often regularized by Bayesian priors or tree parameters (e.g., terminal nodes). Some examples include Bayesian additive regression trees (BART; Hill, 2011), causal random forests (Wager & Athey, 2018) and generalized boosted regression modeling (GBM; McCaffery, Rideway, & Morral, 2004).
3. **Regularized targeted learning:** This approach focuses directly on optimizing the estimation of the treatment effect (i.e., ATE). The constructing of data-adaptive estimators is regulated by Lasso types of regularization. For example, outcome highly adaptive Lasso used in TMLE and C-TMLE framework (Ju, Benkeser, & van der Laan, 2020).

The advantage of a penalized regression is that it will typically converge fast and could work with high dimensional data with less risk of overfitting. However, one of the major drawbacks is that most penalized regressions have a strong functional form assumption (i.e., linear). Ensemble trees and regularized targeted learning relax the fixed functional form assumption. With finite sample size, the regularization may also introduce bias (Chernozhukov et al, 2017; Hahn et al., 2020).

Regularization approaches could also be classified based on different objective functions. For covariate selection, the target variable could include the treatment status and/or outcome variable. For regularization, it could be implemented for propensity score modeling (i.e., GBM; McCaffery, Rideway, & Morral, 2004), high dimensional covariate balancing propensity score (Ning, Peng, & Imai, 2017), outcome modeling (i.e., BART; Hill, 2011), or doubly-robust estimation (Outcome Adaptive Lasso (OAL); Shortreed & Ertefaie , 2017).

2.1.3 Classification of Causal Hybrid Methods

Any combination of covariate selection methods and regularization methods is called a hybrid method. There are two primary approaches to hybrids. The first approach starts with pre-screening the covariate set with explicit covariate selection and then estimating the corresponding model (i.e., propensity score model or response model) using regularization; I refer to this as the explicit hybrid. The second approach embeds covariate selection implicitly into regularization methods; I refer to this as the implicit hybrid. For example, for Collaborative-controlled Lasso (Ju et al., 2019), the covariate is selected for the propensity score model through cross-validation (wrapper method) and Lasso (causal regularization). The complete classification map for causal feature selection approaches is in Figure 2.6.

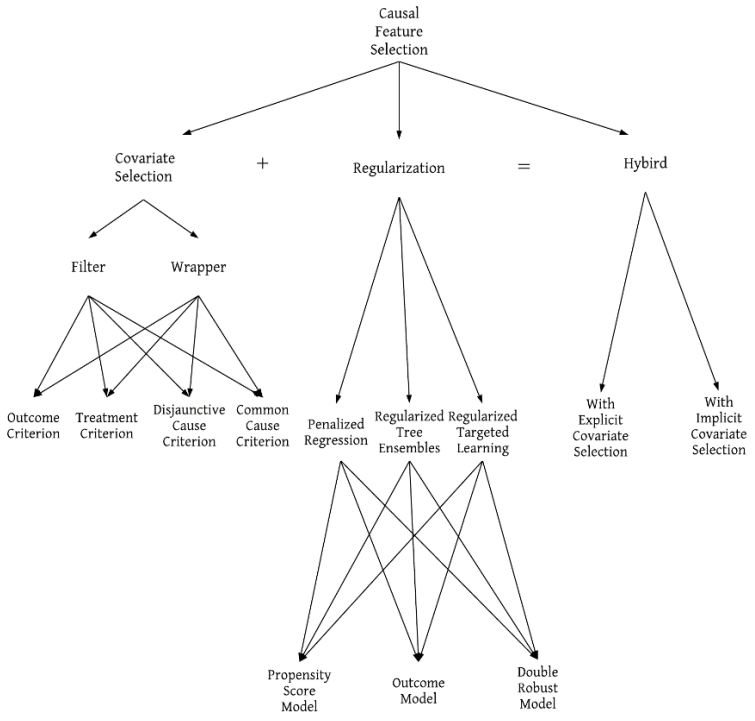


Figure 2.6: Complete classification scheme for causal feature selection.

2.2 Review of Primary Causal Feature Selection Methods

Several additional methods for feature selection have been proposed than can be reviewed within the scope of this research. As a result, I will only provide a detailed description of one exemplar from each of three important categories of causal feature selection methods. These methods were selected because (a) they are widely adapted with good performance in both simulation and real studies, and (b) they are theoretically sound. For causal covariate selection, I will introduce a casual graph-based framework proposed by De Luna, Waernbaum and Richardson (2011) (DWR). For causal regularization, I will present Bayesian Additive Regression Tree (BART) and Generalized Boosted Model (GBM) which are regularized tree-based ensemble methods; and for causal hybrid methods, I will introduce TMLE with Super learning ensemble and C-TMLE.

2.2.1 DeLuna, Waernbaum and Richardson (2011)

De Luna, Waernbaum and Richardson (2011) (DWR) proposed a covariate selection framework for a non-parametric estimation of causal effects. This framework expands on the strong ignorability assumption by connecting it with Pearl's causal graph approach. It aims to identify different types of sufficient covariate sets for confounding control.

Under the Rubin causal model, the strong ignorability assumption is $\{Y^0, Y^1\} \perp T|X$ and $0< P(T|X) < 1$. The sufficient condition for strong ignorability assumption are:

$$\{Y^1 \perp T|X\} \text{ with } 0< P(T = 1|X) < 1, \tag{2.1}$$

$$\text{and } \{Y^0 \perp T|X\} \text{ with } 0< P(T = 0|X) < 1. \tag{2.2}$$

Under the *strong ignorability assumption*, covariate sets that include all covariates that satisfy either (2.1) or (2.2) are sufficient for confounding control. By assuming the distribution over the variable $P(Y^t, T, X)(t = 1,0)$ satisfy *Causal Markov Assumption* (defined in chapter 1) to a causal DAG, the strong relevant covariate could be detected though conditional independence testing.

To clarify the logic behind graph-based covariate selection, I will first formally define Markov blanket as follows: a subset M of X is the Markov blanket of target U, if and only if for any subset $V = X/M$ (where X/M stands for the elements of X that are not in M), $U \perp V \,|M$ (Liu & Motoda, 2008). Suppose Y is the target variable of interest and M is the Markov blanket of Y. Hence, $P(Y|M) = P(Y|M, V = X/M)$ (M is the Markov equivalent class of original covariate set). The extra covariate V could be filtered out if $Y \perp V \,|M$ is satisfied.

Depending on the target variable of interest (i.e., Y or T), it is possible to identify three types of minimum subsets: (1) X_T: include all covariates that predict the treatment (2) X_y include all covariates that predict the outcome (3) Z or Q (depending on the selection algorithm): include only covariates that predict both outcome and treatments (i.e., confounders).

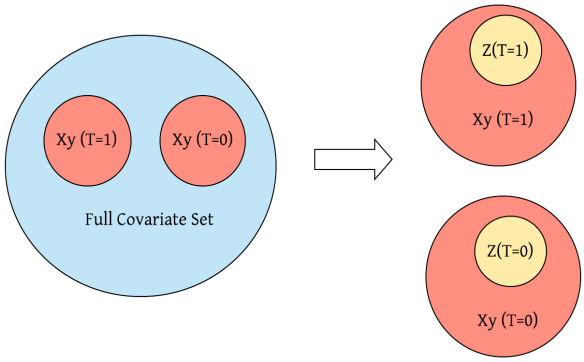


Figure 2.7: DWR algorithm A. It starts with splitting the data set into treated cases and control cases (T=1,0). Within each data set, X_y (minimum covariate set predicting the outcome) is selected by eliminating extra covariates. For each X_y (i.e., $X_{y(T=1)}$ and $X_{y(T=0)}$), a set of confounders $Z_t(T = 1,0)$ is selected (Persson et al., 2017).

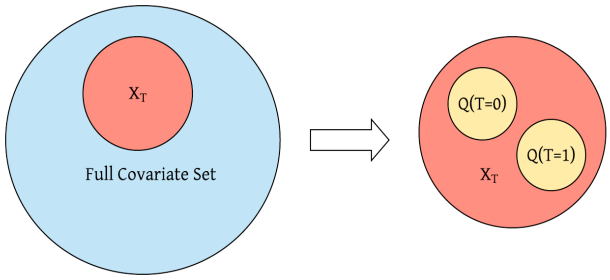


Figure 2.8: DWR algorithm B. It starts with selecting the covariate predicting treatment (X_T). Then, the data set is split into a treated case and a control case; Within each data set, confounders among X_T are selected for a treated case (Q(T=1)) and a control case (Q(T=0)) (Persson et al., 2017).

Two algorithms DWR A (Figure 2.7) and DWR B (Figure 2.8) are operationalized to identify the minimum subset $X_{min} \subseteq X$ such that $\{Y^1, Y^0\} \perp T|X_{min}$ iteratively. DWR A begins with identifying X_y and ends with the minimum confounding set Z. DWR B starts with X_T and ends with Q. Although both Q and Z are the minimum confounding set, they may not be the same.

This frame could end with different target sets that are sufficient for confounding adjustment, such as a set that satisfies the treatment criterion, which is the first step of X_T at algorithm B; a set satisfies the outcome criterion which is $X_y = X_y(T = 1) \cup X_y(T = 0)$ at the first step of algorithm A; a set satisfies the disjunctive cause criterion denoted as $X_{union} = X_y \cup X_T$, and a set that satisfies the common cause criterion which is $X_{min} = Q$ or Z, where $Q = Q(T = 1) \cup Q(T = 0)$ and $Z = Z(T = 1) \cup Z(T = 0)$.

2.2.1.1 Feature Selection and DWR Conditional Independence Tests

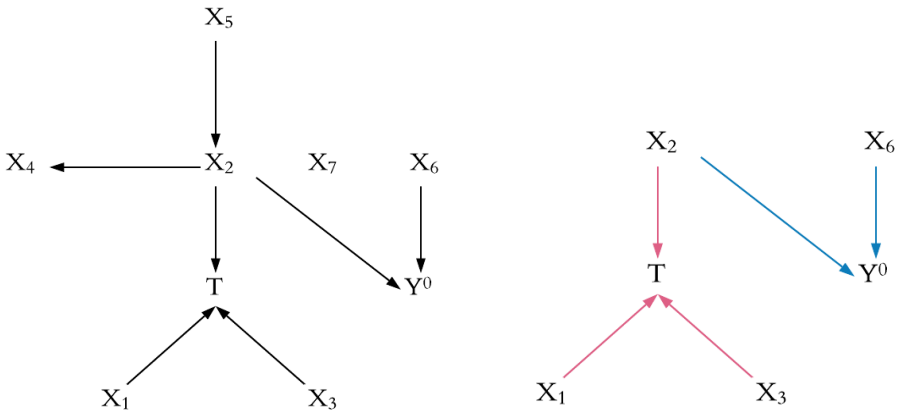


Figure 2.9: A full DAG (left) and its Markov equivalence class (right). Y^1 has the same DAG structure as Y^0. Red outlines the Markov blanket for T and blue outlines the Markov blanket for Y^0.

The feature selection under DWR is accomplished by identifying the minimum Markov Equivalent class of the original DAG. For example, in Figure 2.9, by eliminating irrelevant and weakly relevant covariates, the number of covariates is reduced from 7 to 4. There are various covariates selection methods proposed under DWR.

In general, there are two broad categories that have been proposed. The first category is Markov Blanket learning by using Bayesian networks. For example, Häggström (2018) implemented the Maximum Minimum Parents and Children (MMPC) algorithm and the Maximum Minimum Hill Climbing (MMHC) algorithm for feature selection. Both these algorithms use mutual information to do conditional independence testing to identify relevant features.

There are two Bayesian net searching strategies that aim to identify only strong relevant features by using conditional independence tests: (1) either placing a condition on the entire covariate set (called Markov Blanket discovery) such as Incremental Association Markov Blanket (IAMB) or (2) placing a condition on subsets of covariates (called parents and children discovery) such as MMPC and MMHC. Yu et al. (2019) carried out a simulation study to compare the performances

of local causal discovery algorithms and the result showed that IAMB is the fastest since MMPC has three selection phrases and MMHC has four selection phrases in comparison with IAMB which has only two phrases. IAMB would gain more statistical power in comparison with MMPC/MMHC which has a lower type I error rate, especially for HDLS (high dimensional lower sample size) data.

I incorporated IAMB into DWR X_y framework. The IAMB algorithm consists of forward and backward processes. At the forward phase, each covariate that belongs to the Markov Blanket of the target variable and possibly more are selected. This procedure stops until the selected covariate set does not change. At the backward phase, the false positives are identified, removed, and returned to the final selected sets (Tsamardinos & Aliferis, 2003). A mutual information based conditional independence test is often implemented for identifying strong relevant features. To deal with nonlinearity, continuous variables are discretized based on the 1st and 3rd quantiles.

The second category of methods identifies relevant covariates though measures of variable importance generated by methods such as random forests. Relevant covariates could be filtered out through either importance cut-off values (Häggström, 2018; Persson et al., 2017) or permutation-based hypothesis testing (Keller, 2020). For the latter case, Keller (2020) developed and studied a method for conditional independence testing based on the importance of a random forest variable under permutation and compared it with the importance of a traditional random forest variable.

2.2.2 Targeted Learning

Targeted maximum likelihood estimation (TMLE) is a targeted semi-parametric double robust plug in estimator. Unlike most maximum likelihood estimation (MLE) methods, which minimize the global measure of the model (i.e., mean square error, MSE), TMLE focuses on the target

parameter of interest (i.e., ATE) in a way that both reduces bias and maintains lower estimation variance (van der Laan & Rubin, 2006; van der Laan & Rose, 2011). The observed data set in a TMLE framework is written as: $O(X, T, Y) \sim P(O) = P(X, T, Y)$. The likelihood of the observed data could be orthogonally factorized as:

$$P(O) = P(Y|T, X)P(T|X)P(X). \tag{2.3}$$

Under this decomposition, by assuming SUTVA and *strong ignorability* hold (1.5 and 1.6), the target causal parameter (ATE) could be written as:

$$ATE = E(Y^1) - E(Y^0) = E_x(E(Y|T = 1, X) - E(Y|T = 0, X)). \tag{2.4}$$

The main concerns for ATE using (2.4) are the misspecification of the functional forms of $E(Y|T = 1, X)$ and $E(Y|T = 0, X)$ as well as the bias-variance tradeoff for estimation. TMLE resolves these issues using a two-step procedure (estimation step and targeting step*)* to directly target the parameters of interest.

The estimation step is initiated through an estimation of the response surface model: $(Q^0(T, X) = E(Y|T, X))$, which is used to generate the initial predictive outcomes $(\overline{Q}^0(T, X) = \widehat{E}(Y|T = t, X), \text{where } t = 1 \text{ or } t = 0)$ (van der Laan & Rose, 2011). Then, the targeting step updates $\overline{Q}^0(T, X)$ with the estimated assignment mechanism $g(T|X) = P(T|X)$ by a "clever" covariate:

$$H(T, X) = \left(\frac{I(T=1)}{g(T=1|X)} - \frac{I(T=0)}{g(T=0|X)}\right), \tag{2.5}$$

and fluctuation parameter ϵ_n by using an updating function:

$$\text{logit}(\overline{Q}^1(T, X)) = \text{logit}\left((\overline{Q}^0(T, X)\right) + \epsilon_n H_n(T, X), \tag{2.6}$$

(van der Laan & Rose, 2011). This iterative updating process for $\overline{Q}^2$ $\overline{Q}^3$ .... $\overline{Q}^m$continues until subsequent steps do not result in any further updates of ϵ (i.e., $\hat{\epsilon}^m \to 0$). Finally, the latest estimate

$Q^*(T, X)$ is used to predict pairs of outcomes for both treatment statuses. The ATE are the average difference between these pairs across individuals which can be calculated as:

$$\psi(Q^*) = \frac{1}{n} \sum_{i=1}^{n} \{ \bar{Q}^m(1, X_i) - \bar{Q}^m(0, X_i) \}, \quad (2.7)$$

where i indicates individual observations. This estimation results in an asymptotically unbiased and double robust estimation of the ATE (van der Laan & Rubin, 2006).

TMLE is often implemented with a super learning algorithm for estimation of g and/or Q. The super learning is an ensemble machine learning method based on a generalization of a collection of pre-defined estimation algorithms for prediction (e.g., Lasso, random forest) (van der Laan, Polley, &Hubbard, 2007; Polley & van der Laan, 2010). This builds a weighted combination of estimators, where weights are optimized based on a specific loss-function with cross-validation to guarantee the best overall fit. Super learning allows researchers to use multiple algorithms to outperform a single algorithm in non-parametric statistical models (Polley & van der Laan, 2010).

2.2.2.1 Feature Selection and C-TMLE

While TMLE proved to be successful for high dimensional causal estimation, there are still considerations to be made. One problem of TMLE is overfitting (Zheng & van der Laan, 2011). Collaborative target maximum likelihood estimation (C-TMLE) is an extension of TMLE that adds a stepwise covariate selection procedure for $g^0(T|X)$ and then cross-validates outcomes for a series of estimations of ψ (van der Laan & Gruber, 2010). C-TMLE uses the same loss function as TMLE for Q (i.e., the squared loss $\left(Y - Q(T, X)\right)^2$) to determine whether certain covariates should be included in the propensity score model (van der Laan & Rose, 2011). Let P denote the dimension of the covariate set. For each dimension of the covariate set (i.e., P = 1, 2, 3, …, p), C-TMLE estimates every combination of the covariates in the covariate set and acquires the corresponding TMLE estimator $\widehat{Q}^{P=p}$. Among the series $Q^{P=p}$ within the same dimension, optimal

$\widehat{Q}^{*(P=p)}$ which minimizes the square loss is selected. After selecting the Q^* for each dimension, there will be p optimal Q values $\{\widehat{Q}^{*(P=1)}, \widehat{Q}^{*(P=2)}, \ldots, \widehat{Q}^{*(P=p)}\}$. The C-TMLE algorithm then cross-validates each optimal TMLE estimator and selects the optimal one with minimal loss as the final estimation of Q^* (van der Laan & Gruber, 2010).

Greedy searching in C-TMLE is computationally intensive and scalable to many covariates. It searches through each of the variables in the data set that were not selected in previous iterations. A modified version of C-TMLE allows C-TMLE to be scalable to large data. This is achieved by requiring the algorithm to select covariates in a predefined order (logistic order or partial correlation order) that is specified by the user (Ju et al., 2016).

2.2.3 Regularized Tree Ensembles Methods

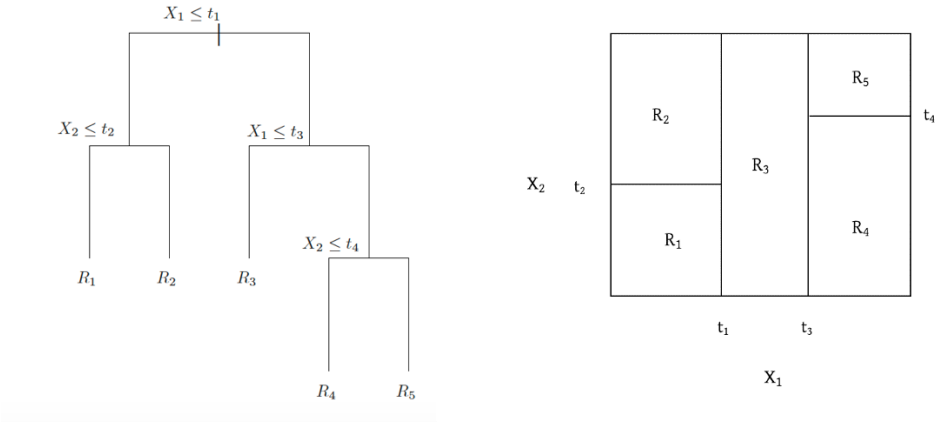


Figure 2.10: A single classification and regression tree (CART).

Tree-based approaches for treatment effect estimation have been widely implemented because of classification and regression tree's (CART) ability to automatically handle nonlinear relationships, including high dimensional interactions that may potentially result in treatment effect heterogeneity.

As shown in Figure 2.10, a single classification and regression tree (CART) split the data into more and more homogenous rectangles denoted by R_i(see the right graph of Figure 2.10). With each of these subsets, the mean for the response variable will be calculated. The splitting point (i.e., t_1) for the tree is determined by minimizing the mean square of the error (MSE) for the regression tree and misclassification error for the classification tree (Hastie, Tibshirani, & Friedman, 2009).

Nevertheless, the CART method can succumb to overfitting. A single CART would predict response surfaces consisting of stepwise jumps between hyperplanes. This will cause relatively unstable prediction performances. The ensemble method could resolve these problems by aggregating multiple weaker learners (single tree), thus resulting in a better ensembled model for prediction. There are several ensemble methods for CART, such as boosting (Freund & Schapire, 1997) and bagging (Breiman, 1996).

Typically, regularization is also added for ensemble methods to reduce the variability of effect estimation. In causal inference, two representative methods for regularized tree-based ensemble are Bayesian additive regression trees (BART) (Chipman, George & McCullagh, 2007; 2010) and Generalized boosted modelling (GBM).

2.2.3.1 Bayesian Additive Regression Tree (BART)

As a high dimension regression method, the Bayesian Additive Regression Tree (BART) was proposed by Chipman, George and McCulloch (2007; 2010). This method fits complex response surfaces as a sequence of prior regulated weak learners. The BART has two parts: a sum of the tree and regularization priors.

$$Y = g(z, x; T_1, M_1) + g(z, x; T_2, M_2) + \cdots + g(z, x; T_m, M_m) + \epsilon, \tag{2.8}$$

$$Y = f(z, x) + \epsilon \text{ and } \epsilon \sim N(0, \sigma^2), \tag{2.9}$$

where the $g(T_i, M_i)$ denotes a single tree i (T is the single tree nodes, z is the treatment status, M is the terminal nodes and $f(z, x)$ is the sum of the tree). T_i, M_i, and σ are treated as parameters and regulated by a pre-specified prior. The fitting process of the BART is similar to a boosting algorithm, but in a stochastic way with Bayesian Makov Chain Monte Carlo (MCMC). After each iteration of MCMC, a new $f(z, x)$ and σ^2 are drawn from its posterior distribution. BART stochastically searches for the best fit of $f(z, x)$ with each single tree to explain the portion of f that is not covered by other trees.

Given the nature of this algorithm, BART could be used in high dimensional settings. It automatically re-weights each covariate based on its ability to explain variation in the outcome, Y. The relative importance of each covariate could be measured as the frequency of each covariate in its sum of tree iterations (Chipman, George, & McCulloch, 2010).

BART can be used to estimate the treatment with a small adjustment. As mentioned in the first chapter, a causal inference problem is also a missing data problem. By placing a condition on the covariate set X, BART could model two response surfaces (treatment and control):

$$E(Y(1)|X = x) = u_1(X), \tag{2.10}$$

$$E(Y(0)|X = x) = u_0(X). \tag{2.11}$$

Both may be used to estimate individual treatment effects. The estimator:

$$CATE = E(Y(1)|X = x) - E(Y(0)|X = x), \tag{2.12}$$

is called conditional average treatment effect (CATE) (Hill, 2011). By averaging over covariate space X, the average treatment effect is:

$$ATE = E_x(E(Y(1)|X = x) - E(Y(0)|X = x)). \tag{2.13}$$

2.3.1.1 Feature Selection and BART Priors

BART regulates the structure of the tree by imposing priors over all parameters of the sum of the tree. These parameters are (T, M) and σ. For the σ prior, BART uses the inverse chi-square distribution to restrict the size of σ to be small (Chipman, George, & McCulloch, 2010). This is the reason why BART estimation has a small standard error in comparison to other approaches. The feature selection is mainly controlled by a prior on T. There are three ingredients for a tree structure prior T. The first part of the probability is that a node at depth d is non-terminal. That is $(\alpha + d)^{-\beta}$, where $\alpha \in (0,1)$ and $\beta \in [0,\infty)$ (Chipman, George, & McCulloch, 2010). The second component is if a node is not a terminal node, there is a probability of the variable being a splitting point of this node. The third component is that the probability over the possible cut-off point for the variable will be used as a splitting variable. They are both regulated by uniform priors. The prior on M regulates the value of the terminal nodes. $p(u|T)$ is distributed as $N(\mu, \sigma^2)$ where the hyper-parameter μ and σ^2 are set in a data driven way. This restricts the value under each of the terminal nodes to be small (Chipman, George, & McCulloch, 2010). Under the MCMC procedure, essential features that are strong predictors of outcome are incorporated automatically under the regularization of BART priors.

2.2.3.2 Generalized Boosted Modelling (GBM)

McCaffrey, Ridgeway and Morral (2014) implemented GBM for propensity score estimation. The GBM model aims to estimate the functional form of the log odds.

$$g(X) = \log\left(p(X)/(1 - p(X)\right), \tag{2.14}$$

in which $p(X)$ is the propensity score by using a boosted tree. The ensemble tree model builds sequentially as a sum of the individual tree.

$$\hat{g}(X) = \widehat{g_0(X_0)} + \widehat{g_1(X_1)} + \ldots \widehat{g_m(X_{0m})} + \epsilon, \tag{2.15}$$

where g_m indicates a single regression tree at m^{th} iteration. The initial fit of the tree is $\hat{g}_0(X) = \frac{\bar{T}}{1-\bar{T}}$, where $\bar{T}$ is the average treatment indicator for the entire sample (McCaffrey, Ridgeway, & Morral, 2014). Then, the algorithm searches for a small adjustment of the function form $h(X)$ which is a regression tree fit to the residual of previous estimation to a maximum log likelihood function:

$$l(g) = \sum_{i=0}^{n} T_i\, g(X_i) - \log\{1 + \exp\big(g(X_i)\big)\}. \quad (2.16)$$

Next, the $(m-1)^{th}$ model for log odds is updated as $\widehat{g_m(X)} = \hat{g}_{m-1}(X) + \lambda h_{m-1}(X)$ where $\lambda \in (0,1]$ and the boosting procedure continues until meeting the stopping criteria (McCaffrey, Ridgeway, & Morral, 2014).

2.2.3.2.1 Feature Selection and GBM Parameter Turning

As a component of the tree model, features that are strong predictors of a treatment are automatically involved in the GBM estimation procedure. There are parameters that play an important role for feature selection in GBM and prevent overfitting. First, at each iteration only 50% of the data randomly sampled in the original data set are used to build a single tree (McCaffrey, Ridgeway &, Morral, 2014). Second, λ is set to 0.0005 to ensure smoothing of the tree fit which would reduce the variance of the tree estimation. Third, GBM limits a maximum of four splits for each single tree. This purring confirms that the GBM will grow a shallow tree which ensures that only the important covariate is involved. Last, the most important feature is the stopping criteria of a GBM function to identify the number of trees to minimize differences between the two treatment groups as one of the following measurements: the mean of the absolute standardized bias (ASB), the maximum of the ASB, the mean of the Kolmogorov-Smirnov (KS) statistic, and the maximum of the KS statistic (McCaffrey, Ridgeway &, Morral, 2014). These stopping criterion enable GBM to handle relatively high dimensional covariate space.

2.3 Discussion

In Chapter 1, the general issues that have existed in the field of causal feature selection have been listed. In this section, I will discuss the specific problems that have not yet been resolved through a literature review. These problems have inspired the simulation design and new methods proposed in Chapter 3. To initiate the discussion, I will decompose the estimation error of any causal estimator as follows:

$$\text{Estimator} - \text{True Causal effect} = \text{Hidden Bias} + \text{Misspecification Bias} + \text{Noise}. \quad (2.17)$$

The hidden bias typically comes from an unmeasured confounding bias, colliding bias, and measurement error (Zhao, Keele & Small, 2019). The misspecification bias usually comes from: (1) the confounding bias and (2) misspecification of the functional form. The noise ordinarily results from finite sample sizes and idiosyncratic variation (Zhao, Keele, & Small, 2019). In general, a well-designed observational study would control most of the hidden bias. The role of causal feature selection is to control the misspecification bias and finite sample variance, which is also known as bias and variance trade-off in machine learning. However, the remaining problem that has not been resolved by previous research revolves around learning how to choose an appropriate feature selection methods/framework and the subsequent estimation strategies with respect to intrinsic characteristics of the data set (i.e., sample size, dimensionality and rate of noise).

2.3.1 Is covariate selection necessary when feature selection is embedded in causal estimation?

While reviewing literature on this topic, one question comes to mind that has also been discussed by previous researchers (e.g., Greenland, 2008; Hoffman et al., 2008): Is covariate selection necessary when there are machine learning alternatives, such as regularization approaches?

There are drawbacks to covariate selection methods as a pre-processing step. For example, there are no selection strategies proven to be uniformly better than others. Some conventional covariate selection methods (i.e., backward elimination) may even delete important confounders. Another issue is that the variance estimation procedures may be tedious (i.e., using bootstrap, Efron (2014)) due to the fact that after covariate selection, the estimation variance should be adjusted since variance is biased downward (Greenland, 2008). However, for the regularization approach, researchers automatically put all covariates into modelling that selects causal features and accounts for selection uncertainty.

Nevertheless, everything comes at a price, and this does not mean that a regularization approach is always better. More researchers (e.g., Chernozhukov et al., 2016; Hahn, Murry and Carvalho, 2017) identify the possibility of regularization introducing bias that originates from the bias variance trade-off. Covariate selection may still be helpful in terms of reducing the dimension of covariate sets to facilitate the convergence rate for non-parametric estimations and maintaining their finite sample asymptotic properties.

To answer this question, we need to develop research designs and benchmarks to empirically study the performance, pros and cons of existing causal estimation approaches, and explore their ability to handle high dimensional covariate space with different sample sizes.

2.3.2 Which covariate selection method is best for pre-processing?

To choose the best pre-processing method, it is critical for researchers to understand the finite sample performances of causal feature selection approach. The more fruitful research question revolves around how large a data set should be when a researcher chooses a specific method for covariate selection. I cannot provide answers based on previous research since the simulation settings, methods being applied, and estimation strategy are diverse between studies. For example,

under the DWR Framework, when the sample size is small (N = 500), the Häggström (2018) MMPC have a relatively small bias, and variance is in the simulation in Witte & Didelez's (2018) study. However, in Keller's (2020) study, the conclusion is different. These differences are due to the linearity of the response surfaces, the number of noise variables included, and the magnitude of the confounding variables, which were all different under the two studies.

Under a finite sample size, several previous studies also investigate the inference for P > N (dimensionality of covariates sets larger than the number of observations) including Wilson and Reich (2014); Ertefaie, Asgharian and Stephens (2014) and Zigler and Doninici (2014), suggesting that this would hinder the performance of causal covariate selection. To answer this question, there must be a study that explores which feature selection methods are more appropriate for a small, moderate, and large sample with low, medium, and high dimensionality. Such studies would help provide insights on how to improve the performance and accuracy of causal feature selection methods based on characteristics of the data set.

2.3.3 Which combinations of covariate selection method with estimation method is the best?

After covariate selection, the next step is to determine an appropriate inference method for treatment effect estimation. Matching, stratification, and weighting aim to create treatment and control groups that are similar in covariate distributions. With sufficient balance achieved on selected key covariates (i.e., confounders), valid treatment effect could be estimated with a smaller standard error than the balance achieved on the entire covariate space.

When working with small sample sizes, modeling may be introduced to improve efficiency. There are three primary model-based approaches to estimate average treatment effects including: (1) modeling of the assignment mechanism (e.g., via logistic regression), (2) modeling of the response surface (e.g., via linear regression), and (3) doubly robust modeling of both assignment

and response surfaces (i.e., via AIPTW). Most modern causal effect estimation involves some type of semiparametric or nonparametric modeling of (1), (2) or (3).

Several simulation studies, case studies and causal inference competitions have been carried out to evaluate the performances of different estimation strategies under high dimensional settings (i.e., Austin, Grootendorst, & Amderson, 2007; Hill, Weiss, & Zhai, 2011, Dorie et al., 2017, Schuler & Rose, 2017). The key takeaway from these studies is that modeling response surfaces have superior performance in comparison to modeling the assignment mechanism alone. Another important trend is the emergence of ensemble learning frameworks (i.e., super learning) which combine different estimation strategies.

For covariate selection, the question that remains and is worth further exploration is whether particular estimation strategies interact with covariate selection strategies. Furthermore, there is a dearth of evidence as to the finite sample properties of causal effect estimators in conjunction with covariate selection.

Chapter 3: Methodological Extensions and Study Designs

In the previous chapter, I have identified three primary causal feature selection approaches that standout for their empirical performance and theoretical importance: (1) the DWR framework, (2) C-TMLE, and (3) tree-based methods. These methods have the potential to be integrated into future applied research.

DWR as a covariate preprocessing framework (i.e., filter methods) could be incorporated into various causal estimation methods due to its flexibility (i.e., the target covariate set for the DWR algorithm can be adapted to fit propensity score, response model, and doubly robust approaches). Based on previous simulation work with DWR algorithms, the outcome criterion (Xy algorithm) and disjunctive cause criterion ($X_T \cup X_y$) performed well and are worthy of further exploration. Under these two selection algorithms, Bayesian networks with the MMPC algorithm and its discretized version proposed by Häggström (2018) and random forest-based methods proposed by Keller (2020) are promising methods for conditional independence testing.

Targeted learning represents a new field in machine learning that focuses on treatment effect estimation with high dimensions and complex covariate relationships. Estimation methods like C-TMLE automatically involve feature selection and improve TMLE's finite sample performance. Super learning ensembles various causal learning algorithms for further enhancements to estimation accuracy.

Ensembled trees can select relevant covariates automatically and work well with higher-order interaction terms. BART has proven to be promising under high dimensional and heterogeneous treatment effect settings in recent studies and data competitions (Dorie et al, 2017). Furthermore, GBM is widely used as a nonparametric method for high dimensional propensity

score estimation (Parast et al, 2018). However, for these three methods to be adapted into applied research, there are still challenges, as discussed at the end of the second chapter.

The first part of this chapter will extend the framework of DeLuna, Waernbaum and Richardson (2011) into regularization settings. Under the extended framework, the covariate selection methods Adaptive Lasso, Double Lasso and permutation-based XGBoost will be introduced.

To be clear, these methods (Adaptive Lasso, Double Lasso under the DWR framework and XGBoost using permutation-based variable selection) are new extensions and novel applications of these existing methods. In particular, Adaptive Lasso has not been used in the DWR framework for variable selection, likely because DWR requires conditional independence testing, which Adaptive Lasso does not provide. However, in the next subsection, I argue that regularization methods such as Adaptive Lasso can, indeed, be profitably used in the DWR framework. Further, I adapt XGBoost, a method based on ensembles of boosted trees that uses regularization, for variable selection by running it within a permutation testing framework.

The second part of this chapter will introduce the design of the empirical Monte Carlo simulation and a follow up case study that will focus on finite sample performances for casual feature selection with constant treatment effect.

3.1 Methodological Extensions

As discussed in chapter 2, the DWR framework was proposed under causal graphic theory. Through conditional independence tests, the Markov Blanket for the target variable could be identified based on different covariate selection criterion. However, conditional independence-testing based approaches are difficult to implement under nonparametric and in a high dimensional covariate space (Jiao et al., 2015; Paninski, 2003). The minimum subset is not necessarily the best subset for causal estimation. Retaining proxy covariates may avoid certain types of estimation bias

and improve estimation efficiency (VanderWeele, 2019). Instead of using conditional independence tests, DWR framework could be modified for regularization-based covariate selection methods.

3.1.1 Extended X_y Framework

To extend the X_y framework, consideration is given to a simple case in which the statistical relationship between Y and X are linear and T is binary treatment,

$$Y = \tau T + \beta X + E, \tag{3.1}$$

where β is a column vector in R^d and E is a zero mean random variable that is independent of X and T. In terms of the potential outcome framework, $Y^0 = \beta X + E^0$ and $Y^1 = \tau + \beta X + E^1$, where E^0 and E^1 are zero mean random error that is independent of X. In observational studies, it is possible to rewrite the assumptions of DeLuna, Waernbaum and Richardson (2011) as follows.

Assumption 1. Let $E^0 \perp T|X$.

Assumption 2. Let $E^1 \perp T|X$.

Assumption 3. Let $0 < P(T = 1|X) < 1$.

Assumption 4. Let $0 < P(T = 0|X) < 1$.

Sometimes, it is necessary to have stronger version of Assumption 1 and Assumption 2. That is,

Assumption 5. $(E^0, E^1) \perp T|X$.

Under Assumption 1, 2, 3 and 4, there is $\tau = E(Y^1 - Y^0) = E(E(Y|X, T = 1) - E(Y|X, T = 0))$. Under Assumption 1 and Assumption 3, $E(Y|X, T = 0) = \beta X$. To estimate β within control case, there is

$$\widehat{\beta}_{(T=0)} = \text{argmin}_\beta\{\|Y_{(T=0)} - X\beta_{(T=0)}\|\} = \beta_{(T=0)} + X^{-1}E^0. \tag{3.2}$$

Let X^{-1} denote the (Moore-Penrose) pseudoinverse. Due to the finite sample size (e.g., $P > N$), $X^{-1}\,\widehat{E}^{0} \neq 0$, Regularization may be used to recover under assumptions related to the sparsity and conditions of the design matrix as follows:

$$\widehat{\beta}_{(T=0)}^{\text{regularization}} = \text{argmin}_{\beta}\{\lambda_0 \|\beta_{(T=0)}\|_1 + \|Y_{(T=0)} - X\beta_{(T=0)}\|\}, \tag{3.3}$$

where λ_0 is the regularization coefficient which will control for overfitting (i.e., $X^{-1}\,\widehat{E}^{0} \to 0$). Under the strong assumption (i.e., Theorem 1 of Janzing, 2019), there is analogy between confounding and overfitting. Therefore, X_0 which is the subset of X with coefficients $\widehat{\beta}_{(T=0)}^{\text{regularization}} \neq 0$ satisfies $E^0 \perp T|X_0$ and similarly, X_1 which is the subset of X with coefficient $\widehat{\beta}_{(T=1)}^{\text{regularization}} \neq 0$ satisfies $E^1 \perp T|X_1$. Under Assumption 5, there is $(E^0, E^1) \perp T|X_y$, where $X_y = X_1 \cup X_0$.

3.1.1.2 Extended X_z Framework

To extend X_z, (3.1) is rewritten as two equations

$$T = p(T = 1|X) = \frac{1}{1+\exp(-\beta_c X)} + v, \tag{3.4}$$

$$Y = \tau T + \beta_d X + E, \tag{3.5}$$

where β_c and β_d column vector in R^d, E is the zero mean random variable that is independent of X and T; and v is the zero mean random variable that is independent of X. Risk factors are covariates with $\beta_d \neq 0$ and $\beta_c = 0$, confounders are with $\beta_d \neq 0$ and $\beta_c \neq 0$ and instrumental variables are $\beta_d = 0$ and $\beta_c \neq 0$. By assuming $(Y^0, Y^1) \perp T|X$ and $0 < P(T|X) < 1$, $Y = \tau(\frac{1}{1+\exp(-\beta_c X)} + v) + \beta_d X + E$. To estimate T with a limited sample, the error term is $\hat{v} \neq 0$.By adding a regularization term on β_c in equation (3.4), the result is

$$\widehat{\beta_c}^{\text{regularization}} = \text{argmin}_{\beta_c}\{\log(1 + \exp(-\beta_c X)) + \lambda_c \|\beta_c\|_1\}. \tag{3.6}$$

Controlling for overfitting of T will lead to $\hat{v} \rightarrow 0$. Under the strong assumption (i.e., Theorem 1 of Janzing, 2019), there is analogy between confounding and overfitting. The covariate set which satisfying $\widehat{\beta_c}^{\,\text{regularization}} \neq 0$ is the set of instrumental variables and confounders which is X_T in DWR.

3.1.2. Adaptive Lasso

Lasso for covariate selection has been previously used within the DWR framework (Persson et al. 2016). However, previous studies (Fan & Li, 2001; Zou, 2006) indicate that Lasso coefficients provide biased estimates for the coefficients under high dimensional settings. Instead, the adapted version of lasso (i.e., adaptive lasso) could be written as follows:

$$\hat{\beta} = \text{argmin}_{\beta}\{(y - \sum_{j=1}^{p} X_j\beta)^2 + \lambda \sum_{j=1}^{p} \widehat{\omega_j}\beta_j\}, \tag{3.7}$$

where $\omega_j = |\widehat{\beta_j(OLS)}|^{-\gamma}$ such that $\gamma > 0$ for coefficient j. The adaptive lasso has oracle properties for high dimensional covariate selection (Fan & Li, 2006; Zou, 2006). In practice, when covariates are highly correlated, Zou (2006) suggests using cross-validated ridge coefficients as adaptive weights since they are more stable than OLS coefficients. The adaptive lasso could be incorporated into the extended X_y algorithm as a two-stage procedure as follows.

Stage one: fit ridge regression and use the inverse of the ridge coefficients as weights

$$\hat{\beta}_{\text{ridge}} = \text{argmin}_{\beta}\{(y - \sum_{j=1}^{p} X_j\beta_j)^2 + \lambda_{\text{ridge}} \sum_{j=1}^{p} \beta_j^2\} \tag{3.8}$$

$$\widehat{\omega_j} = \frac{1}{\hat{\beta}^{\omega}_{\text{ridge}}} \tag{3.9}$$

Stage two: fit lasso regression by adding extra weight at penalty term

$$\hat{\beta}_{\text{ada\_lasso}} = \text{argmin}_{\beta}\{(y - \sum_{j=1}^{p} X_j\beta_j)^2 + \lambda_{\text{lasso}} \sum_{j=1}^{p} \widehat{\omega_j}\beta_j\} \tag{3.10}$$

Covariate with $\hat{\beta}_{\text{ada\_lasso}} \neq 0$ is selected.

3.13. Double Lasso

Because the X_y algorithm focuses exclusively on conditional relationships with potential outcomes, implementations of X_y selection may omit variables with weak outcome relationships but moderate or strong relationships with selection. The disjunctive cause criterion could reduce the chance of omitting such covariates because it takes the union of X_y and X_Z. I have embedded adaptive lasso in $X_y \cup X_Z$ framework with the name double lasso for covariate selection as follows:

Step one: use adaptive lasso to identify all predictors of the outcome.

- Split the data set based on treatment status
- Within each data set, select the predictors of the outcome
- Combine the selected covariates

Step two: use adaptive lasso to identify all predictors of treatment status.

Step three: covariates selected by either step one or step two are included in the final covariate set.

3.14. Permutation-based Regularized XGBoost

As the number of observations and dimensionality of data sets increase, boosted trees often outperform other approaches in both prediction (e.g., Chen& Guestrin, 2016) and causal effect estimation (e.g., BART) in competitions. Whereas random forests grow "full" trees (i.e., trees with low bias and high variance) in a parallel manner, ensemble boosting methods grow "shallow" trees (i.e., trees with high bias and low variance) and the trees are learned sequentially. Therefore, when there are redundant features or highly correlated features, boosting will pick one feature and use it to fit several trees and discard other features since they may not improve the fitting. However, random forests will randomly select features. Subsequently, correlated features are more likely to be included in many trees, which may lead to bias in measures of variable importance for RF-based approaches in contrast with boosting-based approaches.

Nonetheless, the major problem of applying boosted trees for feature selection is that they may overfit, especially with many noise variables. XGBoost is a scalable machine learning system for tree boosting that address this problem by introducing regularization weights by modifying the loss function as follows:

$$\sum_{i=0}^{n} L\left(y_i, f(X_i)\right) + \Omega\,(f), \tag{3.11}$$

where L is the loss function for individual observation i and Ω is the regularization term. The regularization term for a decision tree 1 ... . M is:

$$\Omega\,(f) = \sum_{m=1}^{M} \gamma T_m + \frac{1}{2}\lambda\|\omega_m\|_2 + \alpha\|\omega_m\|_1\,, \tag{3.12}$$

where γ is the complexity of each leaf, T_m is the number of leaves in decision tree m, λ is an L_2 regularization, and α is an L_1 regularization on the weights of each leaf in a tree. The elastic net type regularization described above has demonstrated good performance for $P > N$ feature selection (Zou & Hastie, 2005). At each iteration of stochastic gradient decent, XGBoost includes both row and column subsampling toward improving the generalization performance (Chen & Guestrin, 2016).

Three variable importance measures often used in XGBoost include the gain score, the cover score, and the frequency. I have adapted the gain score as an importance measure; the gain score is a measure of reduction in mean square error (MSE). The gain score tracks the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model. In this way, the gain score is similar to the OOB score used in RF variable importance. Additionally, I have used a permutation based variable importance to select key covariates where the permutation testing procedure is implemented as described by Altmann, Toloşi, Sander, and Lengauer (2010). This is carried out as follows:

Step one: Fit an XGBoost tree to estimate the baseline variable importance for all

covariates.

Step two: Randomly shuffle the row of the covariates and create copies of original data. sets.

Step three: Run XGBoost on each of these data sets.

Step four: Compare the baseline variable importance with its permuted distribution. Covariates that exceed the (1-α) percentile are retained.

3.2 Study Designs

3.2.1 Data Generation

3.2.1.1 Baseline Covariates

I selected 19871 observations and 908 covariates from the Early Childhood Longitudinal Study, Kindergarten cohort of 1998-1999 (ECLS-K) as baseline covariates. These covariates measure children's development, family information, teacher information, and school experiences during the 1998-1999 school year. This data set also included 34 covariates motivated by Morgan, Frisco, Farkas, and Hibel (2010), who evaluated student exposure to special education services on later mathematical achievement. 10 out of 34 covariates were used to simulate the assignment mechanism and response surface. The rest of these covariates were set as noise variables (i.e., they have no causal relationship with the potential outcomes or the treatment selection). All covariates were standardized with a mean of 0 and variance 1. Simulation replications were created by random sampling of individual observations without replacement.

3.2.1.2 Simulation Procedure

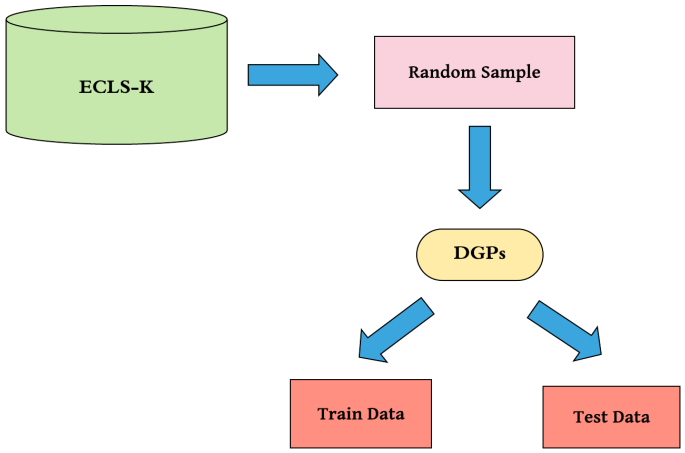


Figure 3.1: Simulation flow chart.

From Figure 3.1, the covariates were first randomly sampled from the original data set. Then, covariates passed though different data generation procedures (DGPs) to create data sets with various simulation factors. These include two broad types of data sets; one was training sets in which pre-processing was performed, and the other was testing sets in which there was an estimation of causal effects both with and without covariate selection. Although simulation results were more accurate with many replications, a number of the covariate selection procedures investigated were extremely computationally expensive. Therefore, to make the simulation tractable, k=100 data sets were used for each scenario. The purpose of this design (i.e., including training and testing data sets) was to reduce the potential for Type I error rate inflation and miscalculation of standard errors related to post-selection statistical inference.

3.2.1.3 Data Generation Procedure

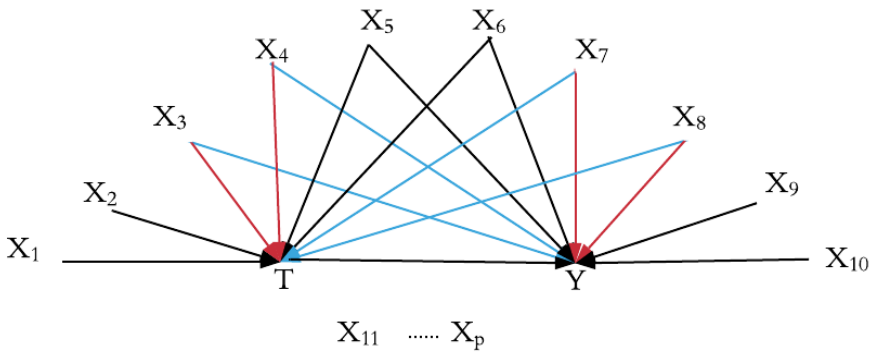


Figure 3.2: Data generation procedure (DGP) when response surface and assignment are linear. Red indicates a strong relationship (with coefficient 1.5) and blue indicates a weak relationship (with coefficient 0.5). Coefficient values are motivated by Shortreed and Ertefaie (2017). X_{11}…. X_p are noise variables.

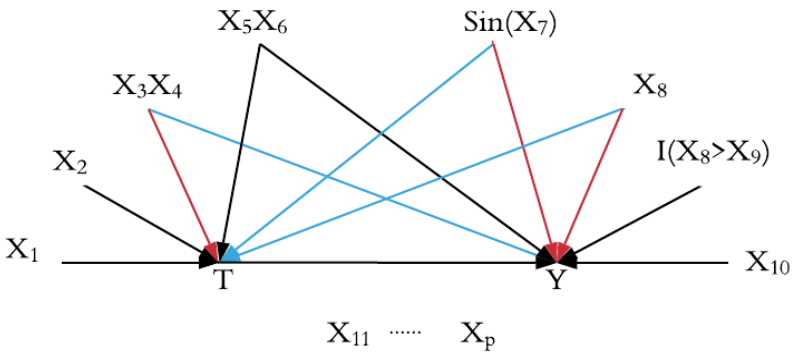


Figure 3.3: Data generation procedure (DGP) when the response surface and assignment are nonlinear. Red indicates a strong relationship (with coefficient 1.5) and blue indicates a weak relationship (with coefficient 0.5). Coefficient values are motivated by Shortreed and Ertefaie (2017). X_{11}…. X_p are noise variables.

Under the strong ignorability assumption, $P(Y_i^1, Y_i^0, T_i|X_i) = P(Y_i^1, Y_i^0|T_i, X_i)P(T_i|X_i)$ where $P(Y_i^1, Y_i^0|T_i, X_i)$ is the response surface model and $P(T_i|X_i)$ is the assignment model. Ten key covariates determine the functional form of $P(Y_i^1, Y_i^0|T_i, X_i)$ and $P(T_i|X_i)$. These variables are:

$X_{treatment} = (X_1, X_2)$ (i.e., covariates that only predict the treatment) , $X_{confounder} = (X_3, X_4, X_5, X_6, X_7, X_8)$ (i.e., covariates predict both treatment and outcome) and $X_{response} = (X_9, X_{10})$ (i.e., covariates only predict the outcome). The rest of the variables $X_{noise} = (X_{11}, \dots, X_p)$ are noise variables. I have simulated different data sets by changing simulation factors that impact either $P(Y_i^1, Y_i^0 | T_i, X_i)$ or $P(T_i|X_i)$. Data are generated as follows:

$$\pi(T_i = 1|X_i) = (1 + \exp\{e_i(X_1, \dots, X_8)\})^{-1}, \quad (3.13)$$

$$T_i \sim \text{Bernoulli}(\pi(T_i = 1|X_i)) , \quad (3.14)$$

$$Y^0{}_{ji} = u_{ji} + \epsilon_{ji}, \quad (3.15)$$

$$Y_{ji}^1 = \tau + u_{ji} + \epsilon_{ji}, \quad (3.16)$$

$$Y_{ji} = Y_{ji}^1 T_{ji} + Y_{ji}^0 (1 - T_{ji}), \quad (3.17)$$

where j indexes a different functional form for individual i, $\tau = 5$ and $\epsilon \sim N(0,1)$. The absolute value of τ is set close to the estimated treatment effect by Morgan et al. (2010). There are two functional forms for the propensity score model (3.13) that is either linear-additive (Figure 3.2):

$$e_1(X_1, \dots, X_9) = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 + \alpha_7 X_7 + \alpha_8 X_8 \quad (3.18)$$

or nonlinear-non-additive (Figure 3.3):

$$e_2(x_1, \dots, x_9) = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 X_4 + \alpha_4 X_5 X_6 + \alpha_5 \sin(X_7) + \alpha_6 X_8 . \quad (3.19)$$

The coefficients for e_1 are set as $(\alpha_1, \dots, \alpha_{10})$ = (1.0, 1.0, 1.5, 1.5, 1.0, 1.0, 0.5, 0.5) and e_2 is set as $(\alpha_1, \dots, \alpha_{10})$ = (1.0, 1.0, 1.5, 1.5, 1.0, 1.0, 0.5, 0.5). These coefficients yield a median value of 0.928 for the proportion of overlapping cases under the linear setting and a median value of 0.990 for the proportion of overlapping cases under the nonlinear setting (also see Figure 3.7). Similarly, for response models there are two functional forms. These are linear and additive:

$$u_1(X_3, \dots, X_{11}) = \beta_1 X_3 + \beta_2 X_4 + \beta_3 X_5 + \beta_4 X_6 + \beta_5 X_7 + \beta_6 X_8 + \beta_7 X_9 + \beta_8 X_{10}, \quad (3.20)$$

and nonlinear-non-additive:

$$u_2(x_3, \ldots, X_{11}) = \beta_1 X_3 X_4 + \beta_2 X_5 X_6 + \beta_3 \sin(X_7) + \beta_4 X_8 + \beta_5 I(X_8 > X_9) + \beta_6 X_{10} . \quad (3.21)$$

The coefficients for u_1 are: $(\beta_1, \ldots, \beta_8) = (0.5, 0.5, 1.0, 1.0, 1.5, 1.5, 1.0, 1.0, 1.0, 1.0)$ and u_2 are: $(\beta_1, \ldots, \beta_8) = (0.5, 1.0, 1.5, 1.5, 1.0, 1.0)$.

The rationale for setting these coefficients is to create three types of confounders. The confounders are Type I = strong relationship with treatment, weak relationship with outcome; Type II = strong relationship with treatment, strong relationship with outcome; and Type III = weak relationship with treatment, strong relationship with outcome. The purpose is to examine the covariate selection methods ability to identify the correct set of covariates for confounding control under a finite sample size and various confounding relationships.

3.2.1.4 Simulation Factors

The simulation factors shown in Table 3.1 are as follows: (1) different sample sizes involving (a) the small sample regime with N= {100, 500} and (b) the moderate sample regime with N= 2000; (2) different functional forms of propensity score models and response models: linear and additive vs. nonlinear and non-additive; and (3) dimensionality of the covariate sets, P= {34, 0.5N, 1.5N}. The dimensionality of the covariate set is varied across three magnitudes. The small case has 34 covariates, as in the original ECLS-K data, the large case has number of covariates equal to half the sample size, and the extreme case has 50% more covariates than the sample size. There are 18 simulation scenarios in total.

Simulation Factors	
Sample size(N)	{100, 500, 2000}
Functional form of T and Y	{(linear, linear), (nonlinear, nonlinear)}
Number of observed covariates (P)	{34, 0.5 × n,1.5 × n}
Factor Design	3×2×3

Table 3.1: Simulation Factors.

3.2.1.5 Measure of the Quality of DGP

The quality of generated data sets is assessed by four measures: degree of nonlinearity, percentage of treated cases, alignment and overlap. These measurements will assess the characteristics of the data set that are important for causal estimation.

Degree of nonlinearity

The measure of nonlinearity of the assignment mechanism is R^2 of regress e on X.

The measure of nonlinearity of the response surface is R^2 of regress Y^1 or Y^0 on X. Among data sets, the degree of nonlinearity ranges from a minimum of 0.31 to a maximum of 0.96 with a median of 0.76 and maximum 0.96.

Percentage of treated

Given that the estimand of interest is the ATE, I try to maintain a 50:50 split for the simulated data. The percentage of treated value is estimated as: $\frac{\mathrm{N_{treated}}}{\mathrm{N}}$. Among data sets, the percentage of treated value ranges from minimum 38% to 70% with a median value of 52%.

Alignment

Kern et al. (2016) define the correspondence between assignment mechanism and the response surface as an alignment. This reflects the severity of the confounding bias. Alignment is measured by the *Pearson correlation* between assignment score π and outcome Y. In this study, alignment value ranges from minimum approximately 0, median approximately about 0.23 and maximum approximately 0.62.

Overlap for the treatment group

Overlap between the treatment and control group is an important factor that impacts the potential for consistent estimation of the ATE. In this simulation, the overlap is measured as the percentage of observations within the interval $[\max(\min(\pi(\mathrm{T}=1)), \min(\pi(\mathrm{T}=0)),$

$\min\left(\max\left(\pi(T=1)\right), \max \pi(T=0)\right)]$, where π is the true propensity score. Under the linear setting, overlap ranges from 0.33 to 0.99, with a median value of 0.93 (see Figures 3.8). Under the nonlinear setting, overlap ranges from 0.63 to 1.00 with a median of 0.99 ((see Figures 3.9).

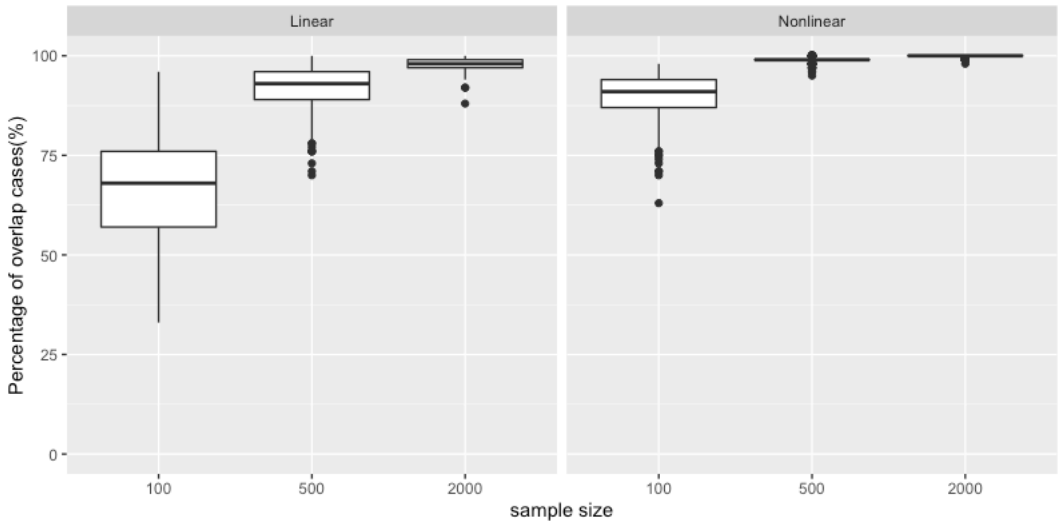
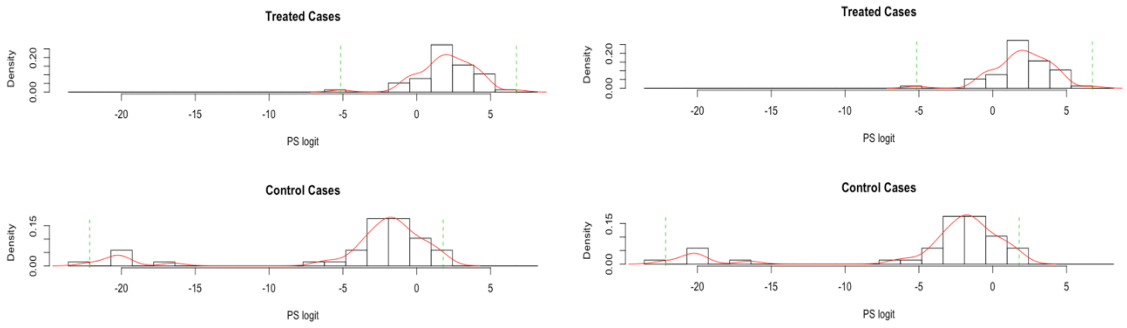
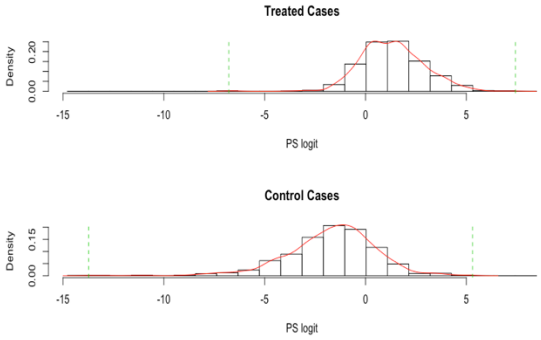


Figure 3.4: The overall distribution of the overlap measurements



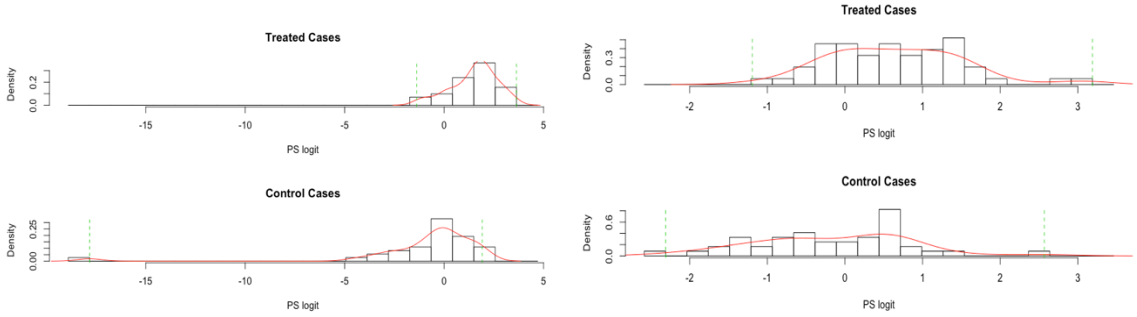
Worst-Case Scenario (0.33)

Moderate Overlap (0.93)

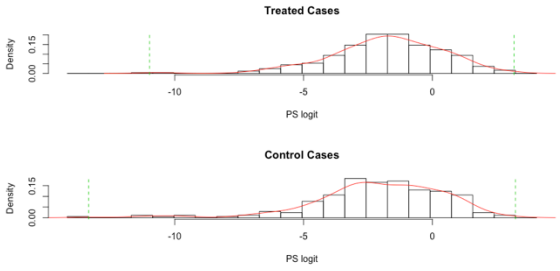


High Overlap (0.99)

Figure 3.5: Examples of overlap in three generated data sets that displayed varying levels of overlap; the response surface and assignment mechanism are linear. The vast majority of generated linear data sets fall somewhere between moderate and high overlap.



Worst-Case Scenario (0.63) Moderate Overlap (0.99)



High Overlap (0.99)

Figure 3.6: Examples of overlap in three generated data sets that displayed varying levels of overlap; the response surface and assignment mechanism are nonlinear. The vast majority of generated nonlinear data sets fall somewhere between moderate and high overlap.

3.2.2 Simulation Studies I, II and III

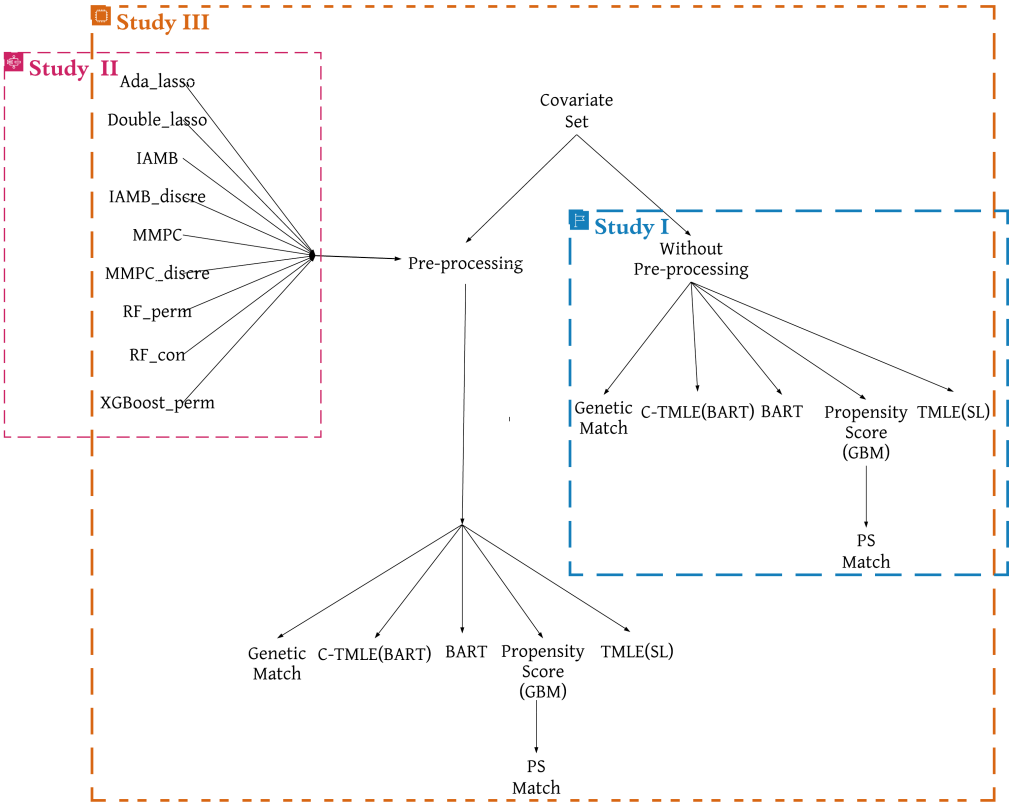


Figure 3.7: Simulation Studies I, II and III.

3.2.2.1 Description of Study I, II and III

Study I: Is covariate selection necessary when feature selection is embedded in causal estimation?

To answer this question, I examine whether pre-processing is helpful or not when there are high dimensional estimation strategies for propensity score and/or response surfaces that select relevant features automatically. Propensity scores are estimated by using Generalized Boosting Modeling

(GBM; Ridgeway et al., 2020) and the treatment effect is estimated by using genetic matching (Genetic Match(PS) with and Genetic Match without propensity score as an extra covariate; Diamond & Sekhon, 2013), propensity score matching (PS Match; Hansen et, al, 2019), BART (BART with the propensity score as an extra covariate), C-TMLE-BART and TMLE-SL.

The design factors for this simulation probe the estimation limits for each strategy under the finite sample size, weak confounding, extreme high dimensionality and complex functional forms for both the response and assignment mechanism. The causal estimand is the overall average treatment effect (ATE). The performances of each method are measured by the average percentage of the absolute bias, average bias, simulation standard deviation and root mean square error.

Study II: Which covariate selection methods are most accurate?

Due to the characteristics of the empirical Monte Carlo simulation, the Markov Blanket for the target variable, the assignment mechanism and response variable are known. The minimum subset set for the confounding control is $\{X_3, X_4, X_5, X_6, X_7, X_8\}$. Since including risk factors X_9 and X_{10} would increase the estimation efficiency, the best method for pre-processing should include all variables in the set $\{X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}\}$ and should avoid both instrumental variables $\{X_1, X_2\}$ and noise variables such as, $X_{11} \ldots . X_P$.

This study aims to evaluate the ability of different covariate selection methods to identify target sets of interest. Three categories of covariate methods including (a) Bayesian net (MMPC, MMPC-discrete, IAMB and IAMB-discrete), (b) regularized regression (Ada-lasso and Double-lasso) and (c) tree-based ensemble (RF-perm, RF-con and XGBoost-perm) are used for covariate pre-processing.

Study III: Do any combinations of covariate selection methods with estimation methods lead to synergistic improvements in estimation accuracy?

Study I explore whether pre-processing is necessary for causal estimation when dimensionality is high. Study II compares different covariate selection methods and evaluates their performances under different simulation settings. In Study III, covariate selection filter methods are crossed with estimation approaches to evaluate performance. The estimates are also compared with ideal (i.e., oracle) selection sets based on outcome criterion, disjunctive cause criterion and estimates without pre-processing to identify the best combination of selection and estimation methods. The measure of performances includes the average percentage of absolute bias, average bias, simulation standard deviation and root mean square error.

3.2.2.2. Implementation Details

MMPC, IAMB, IAMB-discrete and MMPC-discrete run in package *bnlearn* (Scutari, 2010). Both RF-perm and RF-con were implemented by using package *rfvarsel* (Keller, 2020). XGBoost fits were run in package *xgboost* (Chen et al., 2020) and permutation-based importance testing was run through self-written code. Double lasso and adaptive lasso were implemented by using package *glmnet* (Friedman et al., 2020). Package *bartCause* (Dorie, 2020) was used for BART where the propensity score is estimated by BART, also within the package. CTMLE-BART was implemented by using *CTMLE* package (Ju, 2019) where Q was estimated by using "tmle.SL.dbarts2". TMLE-SL was implemented by using *TMLE* package (Gruber, van der Laan & Kennedy, 2020) where the SL library included "SL.glm","tmle.SL.dbarts2", "SL.glmnet" in Q and "SL.glm", "tmle.SL.dbarts.k.5","SL.gam" and "SL.ranger" in G. *twang* (Ridgeway et al., 2020) package was used to estimate the propensity score. *Matching* (Sekhon, 2020) package and *Optmatch* (Hansen et, al, 2019) were used for propensity score and genetic matching.

3.2.2.3 Evaluation Criteria

For variable selection, the performance of each method that was measured by the proportion for which variables are in the target set and by the frequency in which variables are outside of the target set were incorrectly retained. For causal estimations, performance was measured by the average absolute bias, the simulation standard error, and root mean square error have been defined as follows,

$$\text{Average Absolute Percent Bias} = \frac{1}{R}\sum_{r=1}^{R} \frac{|\hat{\tau}_r - \tau|}{\tau} \times 100\%,$$

$$\text{Average Bias} = \frac{1}{R}\sum_{r=1}^{R} \hat{\tau}_r - \tau,$$

$$\text{Simulation Standard Deviation} = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\tau}_r - \bar{\hat{\tau}}_r)^2},$$

$$\text{Root Mean Square Error} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\tau}_r - \tau)^2},$$

where τ is the true ATE, $\hat{\tau}$ is the estimate based on r th replication, and $\bar{\hat{\tau}}_r$ is the average of the estimate over the R total replications.

3.2.3 Empirical Study

Different combinations of covariate selection methods and causal estimation strategies are applied to the ECLS-K data set to estimate the average causal effect of exposure to special education services on math achievement in fifth grade. Since *ignorability* is a strong and untestable assumption, thirty-four relevant covariates recommended by Morgan, Frisco, Farkas, and Hibel (2010) that are either a predictor of the treatment and/or the outcome are included. Morgan et al. (2010) claims they included many additional covariates that only weakly predict the assignment to maximally reduce the potential for selection bias. Therefore, it can be assumed that the *ignorability assumption* holds subjectively and the original covariate set contain subsets in which *ignorability assumption* still holds. There are 429 treated cases and 6933 control cases. Information of the covariates are summarized in Appendix A. The main purpose of this study is to evaluate the

performances of covariate selection methods in causal estimation by incorporating subjective knowledge.

Chapter 4: Results

This chapter presents the results for the simulation studies and empirical data analysis to address the questions raised in chapter 2. In study I, I compare the performances of different causal estimation methods under various sample sizes and dimensionalities. In study II, the focus is on the comparison of different covariate selection methods as pre-processing tools; covariate selection methods are evaluated by the frequency of target variables and the amount of noise variables selected in 100 replications across 18 simulation scenarios. In study III, relevant covariates are selected before causal estimation. The performances of different combinations with covariate selection methods and causal estimation strategies are compared in a simulation study and a closely related empirical study.

4.1 Result of Study I

In general, as expected, smaller sample sizes tend to result in larger SSDs. Similarly, a higher dimensionality of the covariate typically results in larger estimation bias as measured by the AAPB and AB. When dimensionality was extremely high (i.e., P = 1.5N), most of the estimation strategies could not be run due to computational issues. Two exceptions are genetic matching (Gen-Match) and propensity score matching (PS-Match) with sample sizes N=100 and N=500. Most methods were able to be run under the high dimensionality scenario (i.e., when P = 0.5N) with the exception of CTMLE-BART and TMLE-SL when N = 2000, again, due to computational cost.

Under the linear setting with smallest sample size (i.e., N=100) TMLE-SL and PS-Match were associated with the smallest bias. In this case, TMLE-SL was also associated with the lowest simulation standard deviation and, therefore, the lowest RMSE (see the top left cell of Table 4.1). In contrast, BART, CTMLE-BART, Genetic-Match and Genetic-Match (PS) resulted in relatively larger biases. On the other hand, PS-Match was associated with the largest simulation standard

deviation which resulted in the largest RMSE despite having low bias. As expected, performance under BART, CTMLE-BART and PS-Match improved as the sample size was increased. Under different dimensionality conditions (i.e., ranging from P = 34, 50, 150), the trend was similar. When the sample size was increased to 500 and, ultimately, 2000, the differences between BART, CTMLE-BART and TMLE-SL, with respect to bias (AAPB, AB), the simulation standard deviation and RMSE became quite small, with all three methods approximating the RMSE of the oracle method. The SSD of PS-Match continued to diminish with larger samples at a faster rate than other methods, suggesting that with very large sample sizes, PS-Match could be competitive in terms of RMSEs. Genetic-Match and Genetic-Match (PS) were associated with large biases in all conditions. As can be seen in the bottom right-hand cell of Table 4.1, when extreme dimensionality was coupled with the largest sample size condition no methods were computationally feasible.

Under a nonlinear setting with smallest sample size (i.e., N=100), PS-Match performs the best in terms of bias (i.e., AB) with Genetic-Match (with and without PS) and TMLE-SL in second and third place, respectively. The story is much the same as the linear case in that the good performance as measured by bias for PS-Match is outweighed by the very large SSD, which drives up the RMSE. In the nonlinear case, Genetic-Match and TMLE-SL have the smallest RMSEs across the first two dimensionality scenarios (i.e., P = 34 and P = 50). For the extremely high dimensional case with small sample size (i.e., N = 100, P = 150), only PS-Match and Genetic-Match were able to produce estimates; among the two matching methods, Genetic-Match and Genetic-Match (PS) had the lower RMSE across all three dimensionality conditions. As the sample size was increased to N = 500 and 2000, the performance of BART, CTMLE-BART, and TMLE-SL improved markedly with respect to both bias and SSD for the low dimensionality case (P = 34) but not for the high

dimensional case (P = .5N); furthermore, none of these methods were able to produce estimates for the extremely high dimensional scenario (P = 1.5N). As sample size was increased, TMLE-SL did not perform as well as BART or CTMLE-BART with respect to bias and simulation standard error. Furthermore, although the performance of PS-Match, Genetic-Match and Genetic-Match(PS) improved with larger samples, the rate of improvement was much slower here than in the linear case and, in particular, not nearly as fast as the rates of improvement associated with the other approaches (i.e., BART, CTMLE-BART and TMLE-SL).

N=100														
P=34	**AAPB**	**AB**	**SSD**	**RMSE**	**P=50**	**AAPB**	**AB**	**SSD**	**RMSE**	**P=150**	**AAPB**	**AB**	**SSD**	**RMSE**
Oracle	4.12	0.01	0.26	0.26		4.38	0.04	0.27	0.27		3.55	0.01	0.23	0.23
BART	20.65	-1.02	0.58	1.17		24.38	-1.22	0.59	1.36		N/A	N/A	N/A	N/A
CTMLE-BART	24.49	-1.22	0.90	1.52		28.34	-1.42	0.70	1.58		N/A	N/A	N/A	N/A
TMLE-SL	9.52	-0.35	0.49	0.60		10.20	-0.41	0.51	0.65		N/A	N/A	N/A	N/A
PS-Match	33.29	-0.43	2.10	2.13		35.60	-0.07	2.24	2.23		24.86	0.12	1.74	1.68
Gen-Match (PS)	24.70	-1.21	0.69	1.39		25.12	-1.22	0.79	1.45		40.28	-2.01	0.76	2.15
Gen-Match	25.06	-1.23	0.67	1.40		27.26	-1.35	0.65	1.50		31.47	-1.57	0.72	1.73

N=500														
P=34	**AAPB**	**AB**	**SSD**	**RMSE**	**P=250**	**AAPB**	**AB**	**SSD**	**RMSE**	**P=750**	**AAPB**	**AB**	**SSD**	**RMSE**
Oracle	1.98	-0.01	0.12	0.12		1.54	-0.01	0.10	0.10		1.69	0.01	0.11	0.11
BART	2.79	-0.08	0.15	0.17		4.43	-0.21	0.14	0.26		N/A	N/A	N/A	N/A
CTMLE-BART	3.12	-0.10	0.16	0.19		5.19	-0.21	0.22	0.30		N/A	N/A	N/A	N/A
TMLE-SL	2.51	0.02	0.16	0.16		3.43	-0.11	0.18	0.21		N/A	N/A	N/A	N/A
PS-Match	8.36	-0.16	0.50	0.52		8.64	-0.17	0.50	0.52		9.59	0.20	0.56	0.59
Gen-Match (PS)	21.84	-1.09	0.34	1.14		27.91	-1.39	0.39	1.45		N/A	N/A	N/A	N/A
Gen-Match	22.65	-1.13	0.37	1.19		31.44	-1.57	0.36	1.61		45.51	-2.28	0.33	2.30

N=2000														
P=34	**AAPB**	**AB**	**SSD**	**RMSE**	**P=1000**	**AAPB**	**AB**	**SSD**	**RMSE**	**P=3000**	**AAPB**	**AB**	**SSD**	**RMSE**
Oracle	0.98	0.00	0.06	0.06		0.89	0.00	0.06	0.06		0.78	0.00	0.05	0.05
BART	1.30	0.05	0.06	0.08		2.66	-0.13	0.07	0.15		N/A	N/A	N/A	N/A
CTMLE-BART	1.36	-0.04	0.07	0.08		N/A	N/A	N/A	N/A		N/A	N/A	N/A	N/A
TMLE-SL	1.24	-0.01	0.08	0.08		N/A	N/A	N/A	N/A		N/A	N/A	N/A	N/A
PS-Match	7.98	-0.40	0.17	0.43		8.79	-0.44	0.20	0.48		N/A	N/A	N/A	N/A
Gen-Match (PS)	13.70	-0.68	0.20	0.71		N/A	N/A	N/A	N/A		N/A	N/A	N/A	N/A
Gen-Match	22.31	-1.12	0.20	1.13		32.30	-1.58	0.90	1.82		N/A	N/A	N/A	N/A

Table 4.1: Under linear settings, the ATE estimations of Oracle, BART CTMLE-BART, TMLE-SL, PS-Match, Genetic-Match and Genetic-Match (PS). AAPB stands for Average Absolute

Percentage Bias, AB stands for Average Bias, SSD stands for Simulation Standard Deviation and RMSE stand for root mean square error. Oracle is estimated by using the correct functional form of the assignments and response surface.

N=100													
P=34	**AAPB**	**AB**	**SSD**	**RMSE**	**P=50**	**AAPB**	**AB**	**SSD**	**RMSE**	**P=150**	**AAPB**	**AB**	**SSD**	**RMSE**
Oracle	1.68	-0.02	0.11	0.11		1.89	-0.01	0.12	0.12		1.92	0.00	0.12	0.12
BART	34.56	-1.73	0.65	1.85		34.56	-1.73	0.65	1.85		N/A	N/A	N/A	N/A
CTMLE-BART	34.07	-1.70	0.65	1.82		37.66	-1.85	0.90	2.06		N/A	N/A	N/A	N/A
TMLE-SL	28.97	-1.45	0.67	1.59		27.08	-1.35	0.62	1.49		N/A	N/A	N/A	N/A
PS-Match	44.01	-0.95	2.63	2.79		47.29	-1.45	2.82	3.16		44.82	-1.98	1.87	2.68
Gen-Match (PS)	27.38	-1.34	0.83	1.58		24.58	-1.17	0.85	1.45		30.58	-1.51	0.88	1.57
Gen-Match	26.65	-1.32	0.74	1.51		25.60	-1.25	0.76	1.46		31.67	-1.58	0.41	1.64

N=500													
P=34	**AAPB**	**AB**	**SSD**	**RMSE**	**P=250**	**AAPB**	**AB**	**SSD**	**RMSE**	**P=750**	**AAPB**	**AB**	**SSD**	**RMSE**
Oracle	0.68	0.00	0.04	0.04		0.75	0.00	0.05	0.05		0.76	0.00	0.05	0.05
BART	12.66	-0.63	0.19	0.66		23.22	-1.16	0.23	1.18		N/A	N/A	N/A	N/A
CTMLE-BART	12.48	-0.62	0.22	0.66		22.92	-1.15	0.26	1.18		N/A	N/A	N/A	N/A
TMLE-SL	16.56	-0.83	0.22	0.86		24.82	-1.24	0.25	1.27		N/A	N/A	N/A	N/A
PS-Match	19.15	-0.91	0.62	1.10		22.14	-1.08	0.59	1.23		27.93	-1.37	0.67	1.52
Gen-Match (PS)	21.90	-1.09	0.36	1.15		24.33	-1.22	0.39	1.28		N/A	N/A	N/A	N/A
Gen-Match	22.91	-1.15	0.38	1.21		26.43	-1.32	0.39	1.38		31.42	-1.57	0.42	1.63

N=2000													
P=34	**AAPB**	**AB**	**SSD**	**RMSE**	**P=1000**	**AAPB**	**AB**	**SSD**	**RMSE**	**P=3000**	**AAPB**	**AB**	**SSD**	**RMSE**
Oracle	0.35	0.00	0.02	0.02		0.40	0.00	0.02	0.02		0.41	0.00	0.03	0.03
BART	5.39	-0.27	0.08	0.28		22.01	-1.10	0.12	1.11		N/A	N/A	N/A	N/A
CTMLE-BART	5.33	-0.27	0.08	0.28		N/A	N/A	N/A	N/A		N/A	N/A	N/A	N/A
TMLE-SL	6.88	-0.34	0.08	0.35		N/A	N/A	N/A	N/A		N/A	N/A	N/A	N/A
PS-Match	22.45	-1.12	0.19	1.14		24.01	-1.20	0.19	1.21		N/A	N/A	N/A	N/A
Gen-Match (PS)	17.16	-0.86	0.21	0.88		N/A	N/A	N/A	N/A		N/A	N/A	N/A	N/A
Gen-Match	21.19	-1.06	0.19	1.08		30.09	-1.50	0.27	1.53		N/A	N/A	N/A	N/A

Table 4.2: Under nonlinear settings, the ATE estimations of Oracle, BART, CTMLE-BART, TMLE-SL, PS-Match, Genetic-Match and Genetic-Match (PS). AAPB stands for Average Absolute Percentage Bias, AB stands for Average Bias, SSD stands for Simulation Standard Deviation and RMSE stand for root mean square error. Oracle is estimated by using the correct functional form of the assignments and response surface.

4.2 Results of Study II

4.2.1 Small Sample Size

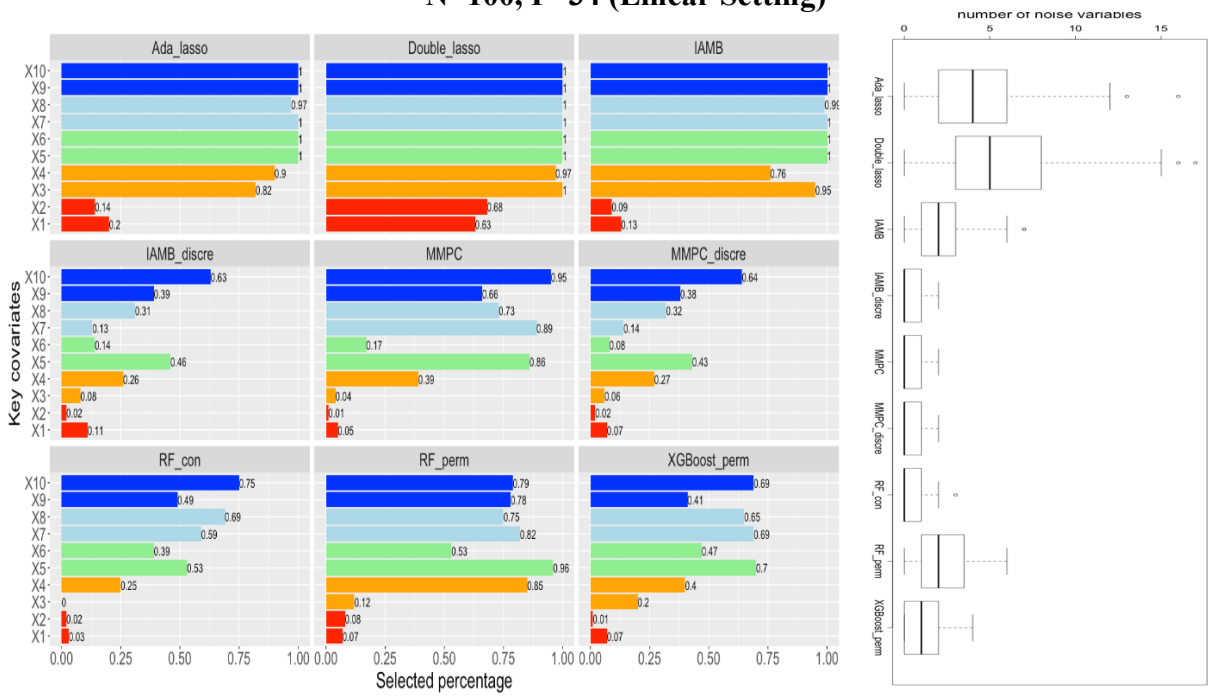


Figure 4.1: Results for the linear case with a small sample size (N = 100) and small dimensionality (P = 34). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 (in red) are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

N=100, P=50 (Linear Setting)

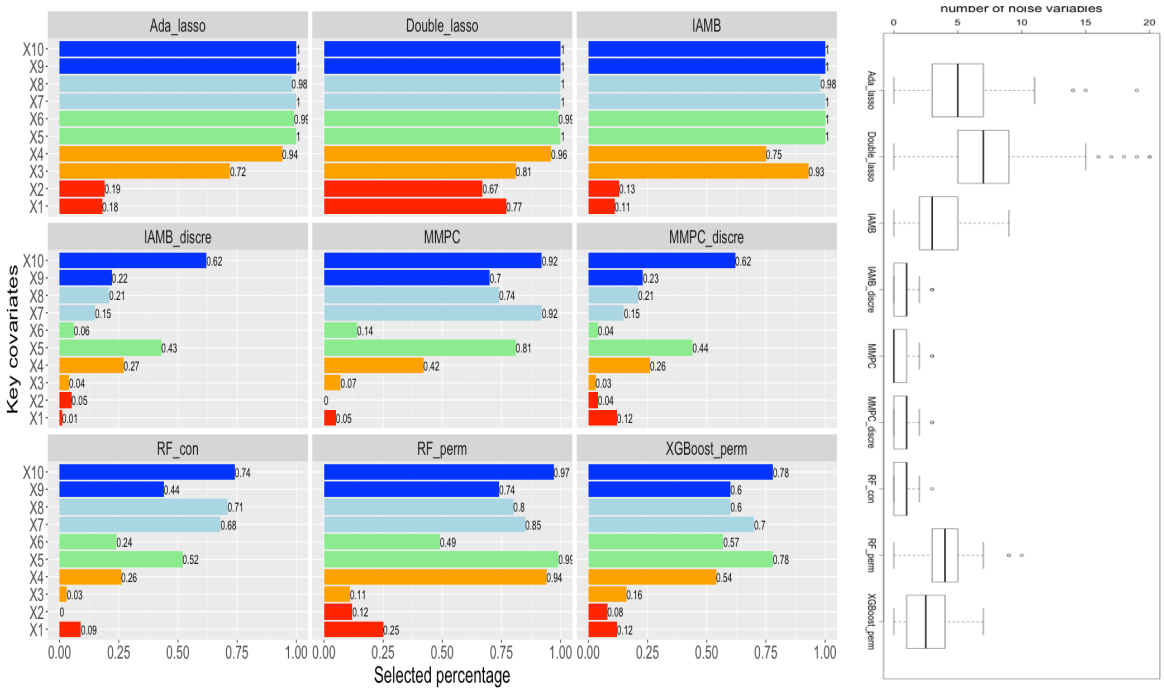


Figure 4.2: Results for the linear case with a small sample size (N = 100) and high dimensionality (P = 50). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 (in red) are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

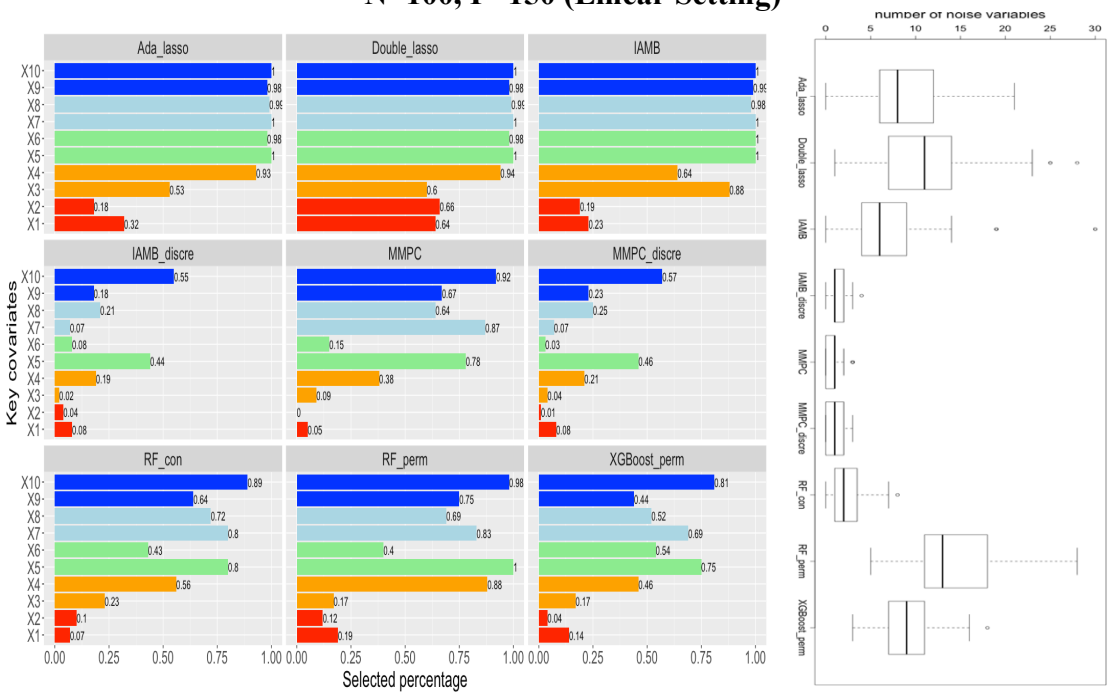


Figure 4.3: Results for the linear case with a small sample size (N = 100) and extreme high dimensionality (P = 150). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 (in red) are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

N=100, P=34 (Nonlinear Setting)

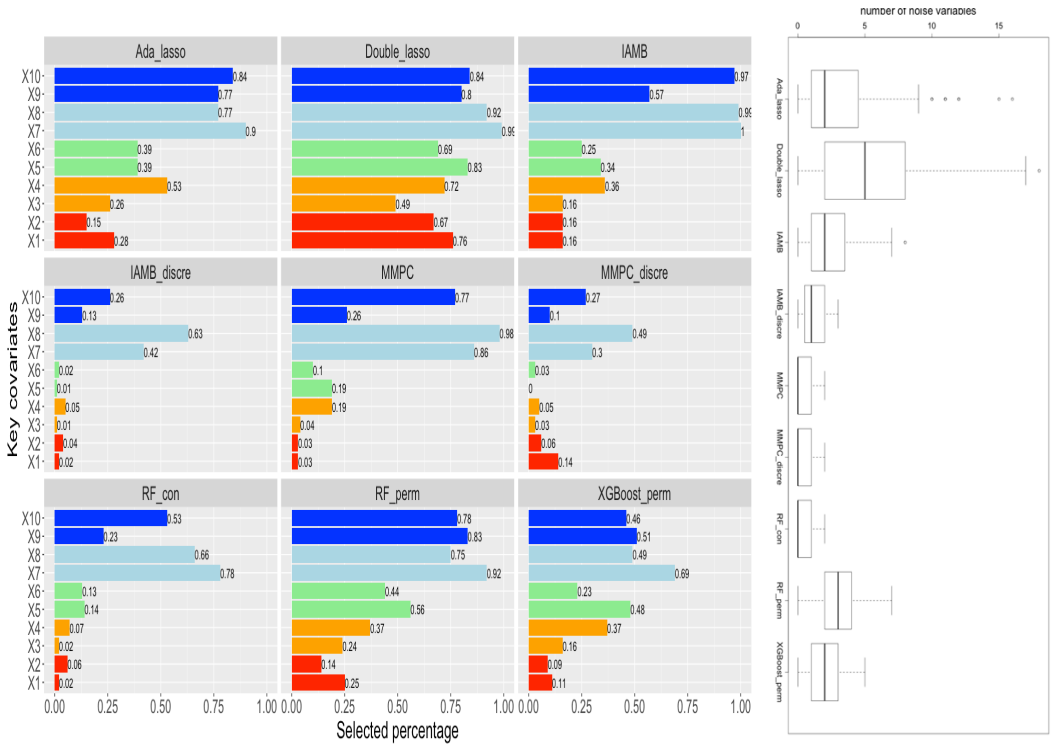


Figure 4.4: Results for the nonlinear case with a small sample size (N = 100) and small dimensionality (P = 34). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 (in red) are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

N=100, P=50 (Nonlinear Setting)

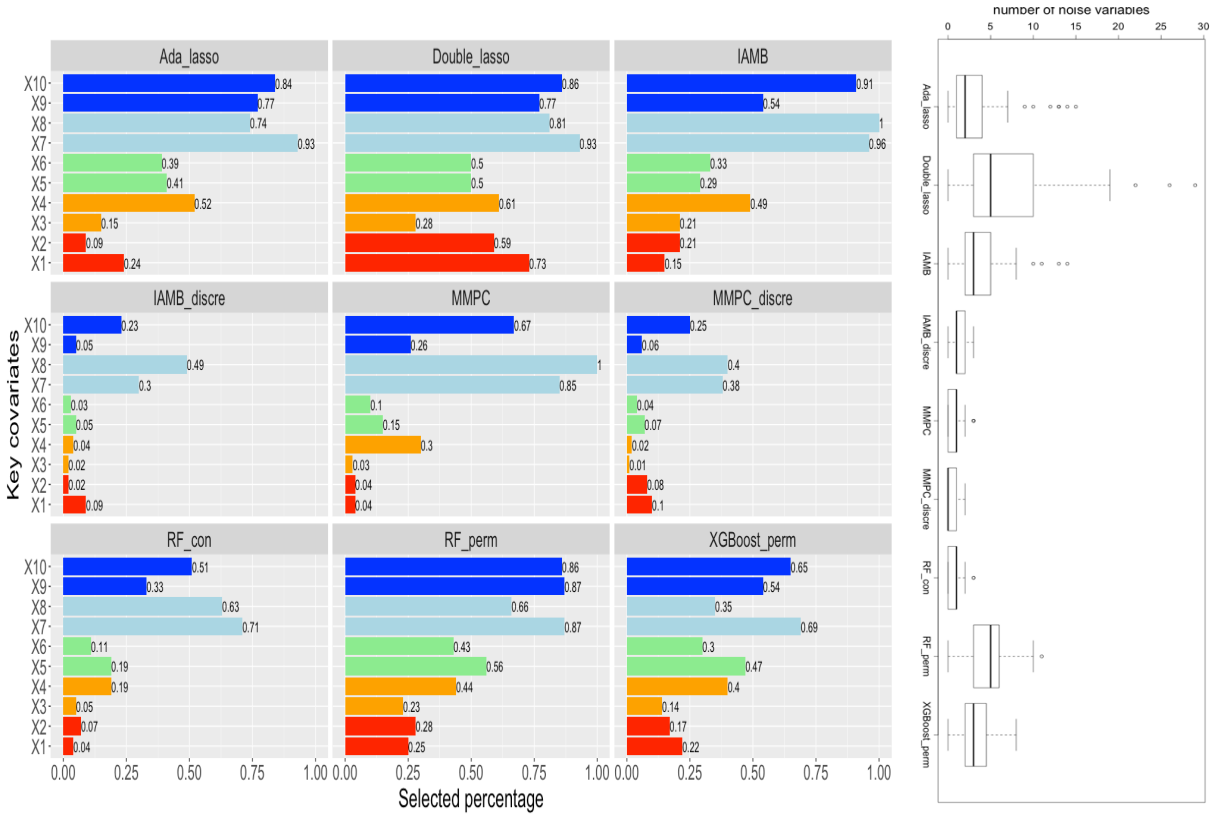


Figure 4.5: Results for the nonlinear case with a small sample size (N = 100) and high dimensionality (P = 50). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 (in red) are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

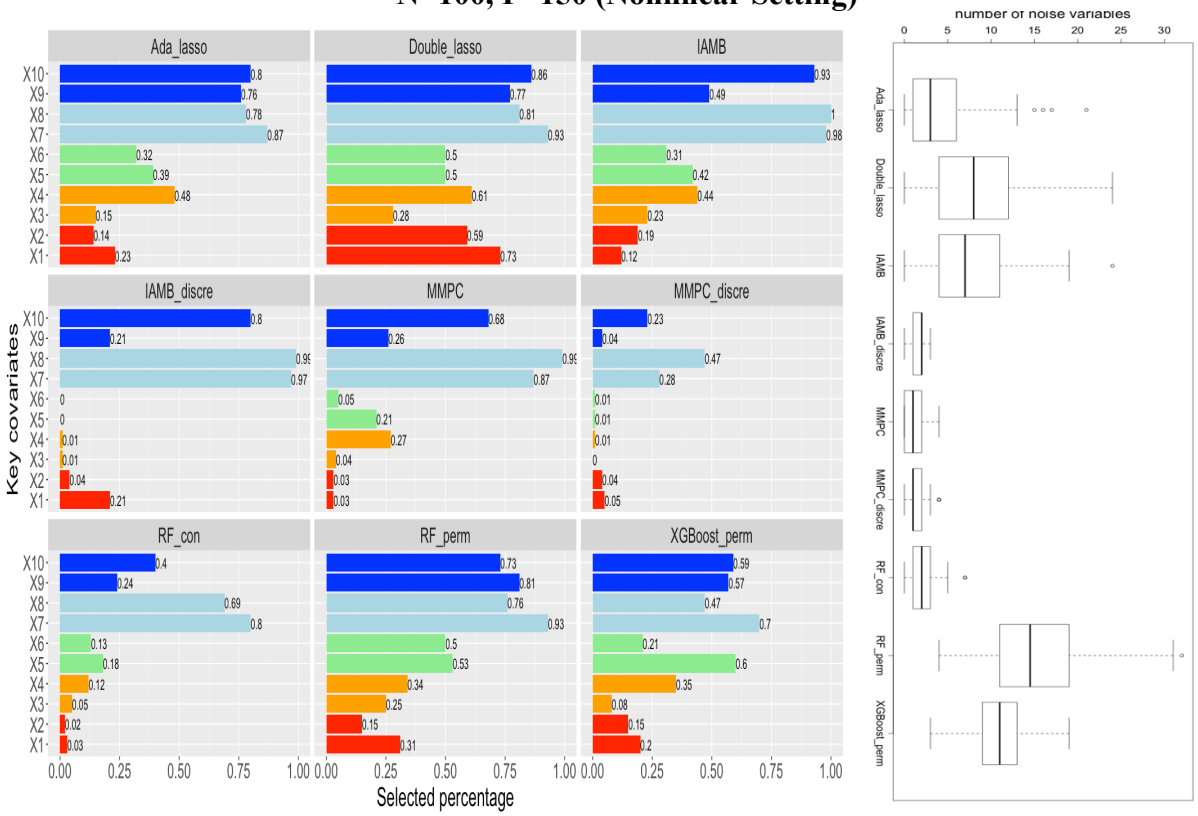


Figure 4.6: Results for the nonlinear case with a small sample size (N = 100) and extreme high dimensionality (P = 150). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 (in red) are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

By comparing the average inclusion rate for confounding variables, risk factors, and instrumental variable, an overview about performances of different selection approaches as pre-processing tools can be provided. Under the small sample (N=100) and linear setting (Figure 4.1-Figure 4.3), IAMB was most accurate, with an average rate of 0.95 for correctly selecting risk factors and confounding variables (analogous to power in this context), and an average rate of 0.14 for incorrectly including instrumental variable. Adaptive lasso was a close second in terms of average inclusion rate for target variables with an inclusion rate of approximately 0.90 for risk factors and confounding variables, however, this approach, had a slightly higher inclusion rate for the instrumental variables at around 0.20. The median number of noise variables selected by IAMB was less than 7 for all three dimensionalities; the same was true for Double Lasso.

Double Lasso is the only method that is set up, by design, to include selection-only predictors (i.e., instrumental variables) because it targets the disjunctive cause criterion. Thus, double lasso was much more likely than other methods to select instrumental variables, with an average rate of 0.68 for X_1 and X_2. Furthermore, the double lasso algorithmically provides two chances to pick up confounding variables (i.e., on the outcome side and on the assignment side), and this was reflected in its higher average rate of selection of confounding and risk factors (0.97). Tree-based methods (i.e., RF-con, RF-perm and permutation-based regularized XGBoost) and discretized Bayesian nets (i.e., MMPC-discre and IAMB-discre) were more likely to omit key covariates under this setting. The average inclusion rates for key covariates (i.e., confounders and risk factors) are all below 0.8 for these methods.

Under a small sample (N=100) and in a nonlinear setting (Figure 4.4-Figure 4.6), the performance of most of the methods is not good. The largest average inclusion rate of key covariates comes with Double Lasso (about 0.70). Double Lasso also has an average rate of

approximately 0.68 in including the instrumental variable. Permutation based random forest has an average rate of approximately 0.62 in including the key covariates, and an average rate of approximately 0.23 in including the instrumental variables. That is the best combination under this setting. Adaptive Lasso is a close second with approximate average rate of 0.59 in including the key covariates and average rate of 0.18 in including the instrumental variables. Bayesian net-based methods face difficulties in detecting interaction terms. permutation-based regularized XGBoost and RF-con have a high average rate of omitting the key covariates. Under both scenarios, an increase of the dimensionality results in a high rate for including noise variables and/or instrumental variables for RF-perm, IAMB and Double Lasso. However, due to a small sample size, the number of noise variables selected is not large.

4.2.2 Medium Sample Size

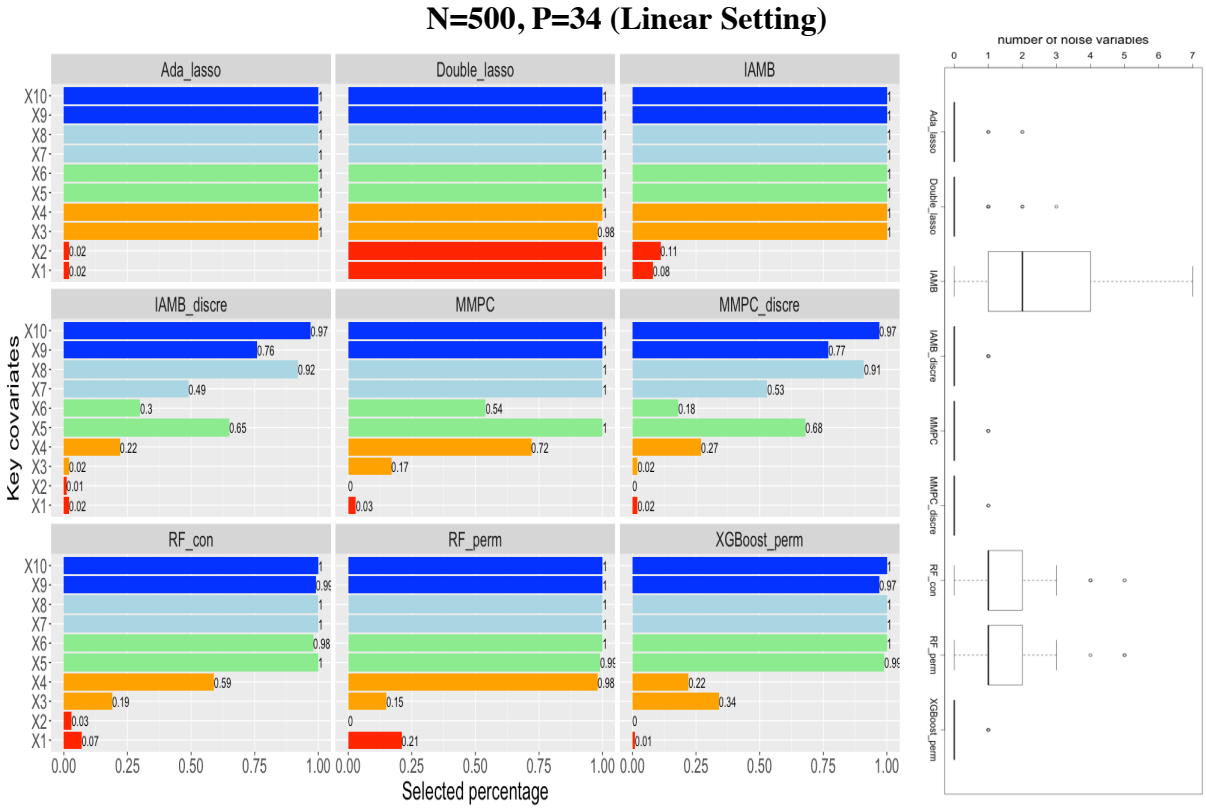


Figure 4.7: Results for the linear case with a medium sample size (N = 500) and small dimensionality (P = 150). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 , in red, are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

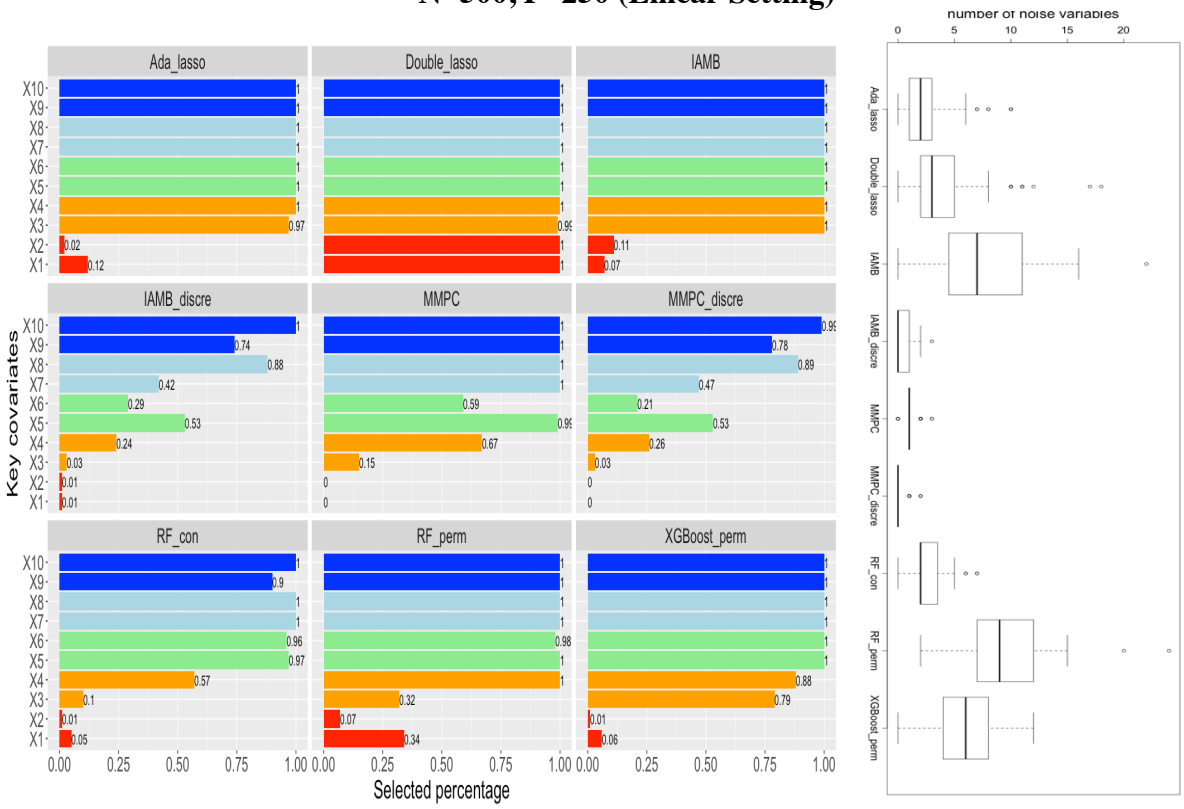


Figure 4.8: Results for the linear case with a medium sample size (N = 500) and high dimensionality (P = 250). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 , in red, are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

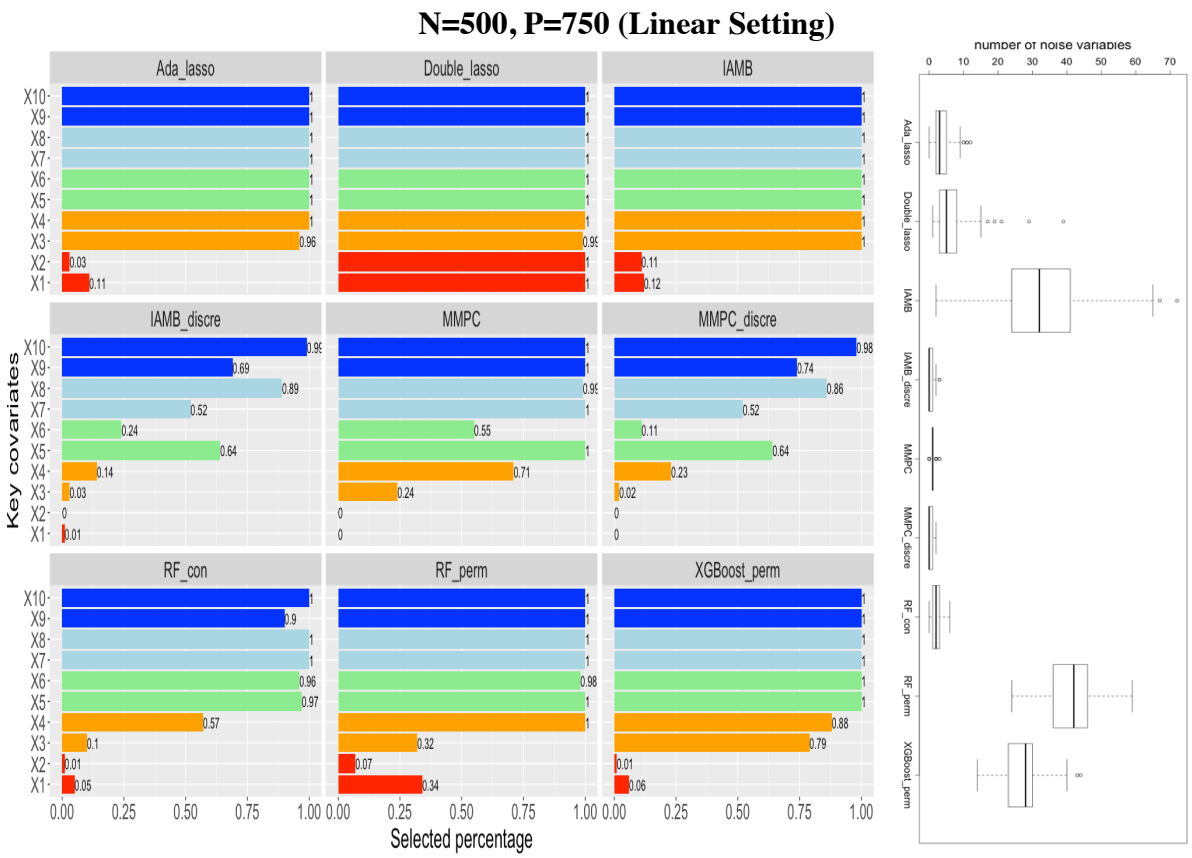


Figure 4.9: Results for the linear case with a medium sample size (N = 500) and extreme high dimensionality (P = 750). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 , in red, are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

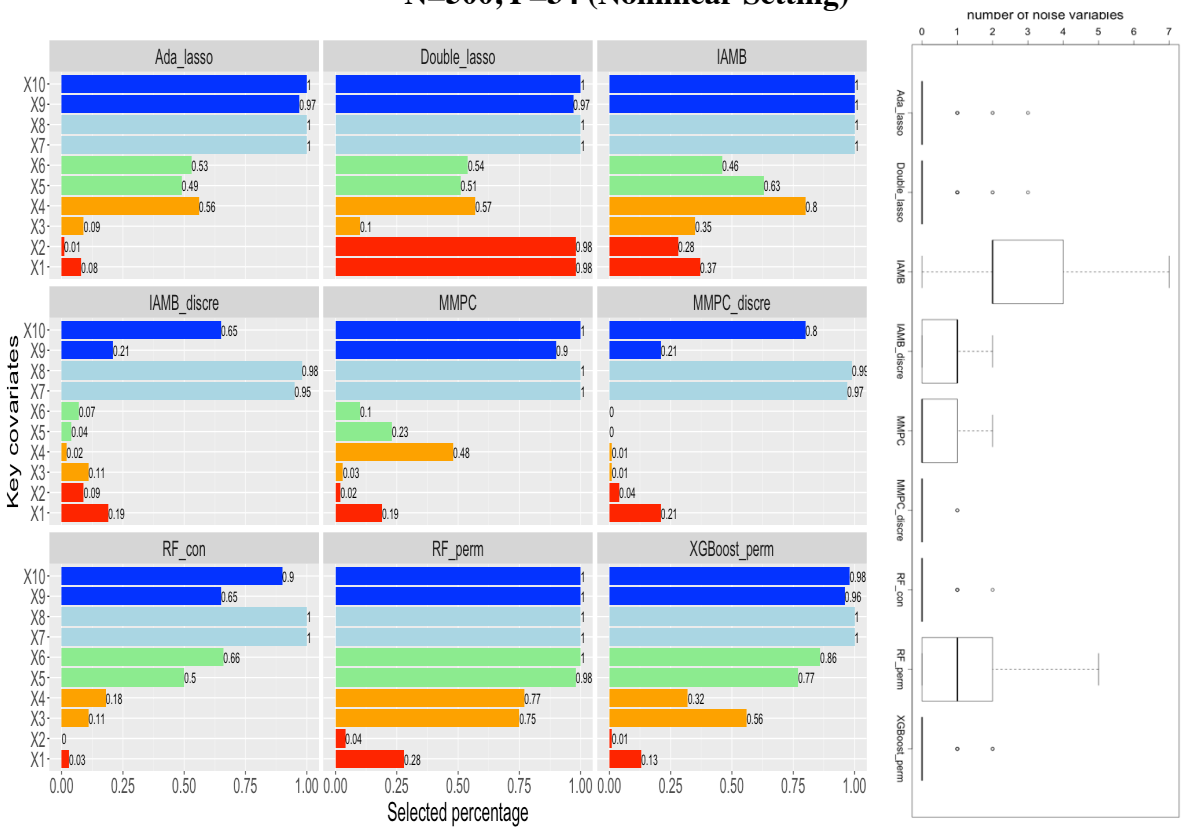


Figure 4.10: Results for the nonlinear case with a medium sample size (N = 500) and low dimensionality (P = 34). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 , in red, are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

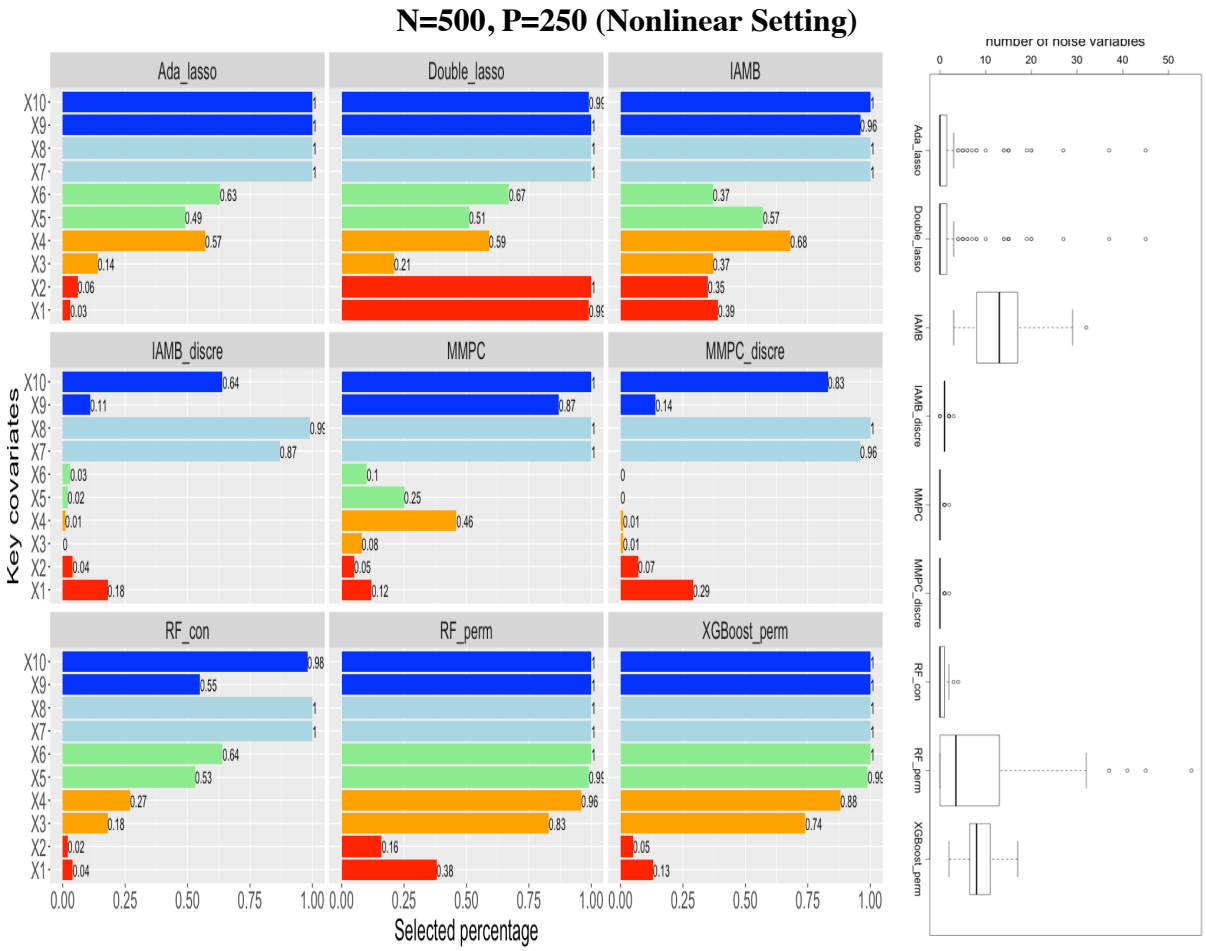


Figure 4.11: Results for the nonlinear case with a medium sample size (N = 500) and high dimensionality (P = 250). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 , in red, are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

N=500, P=750 (Nonlinear Setting)

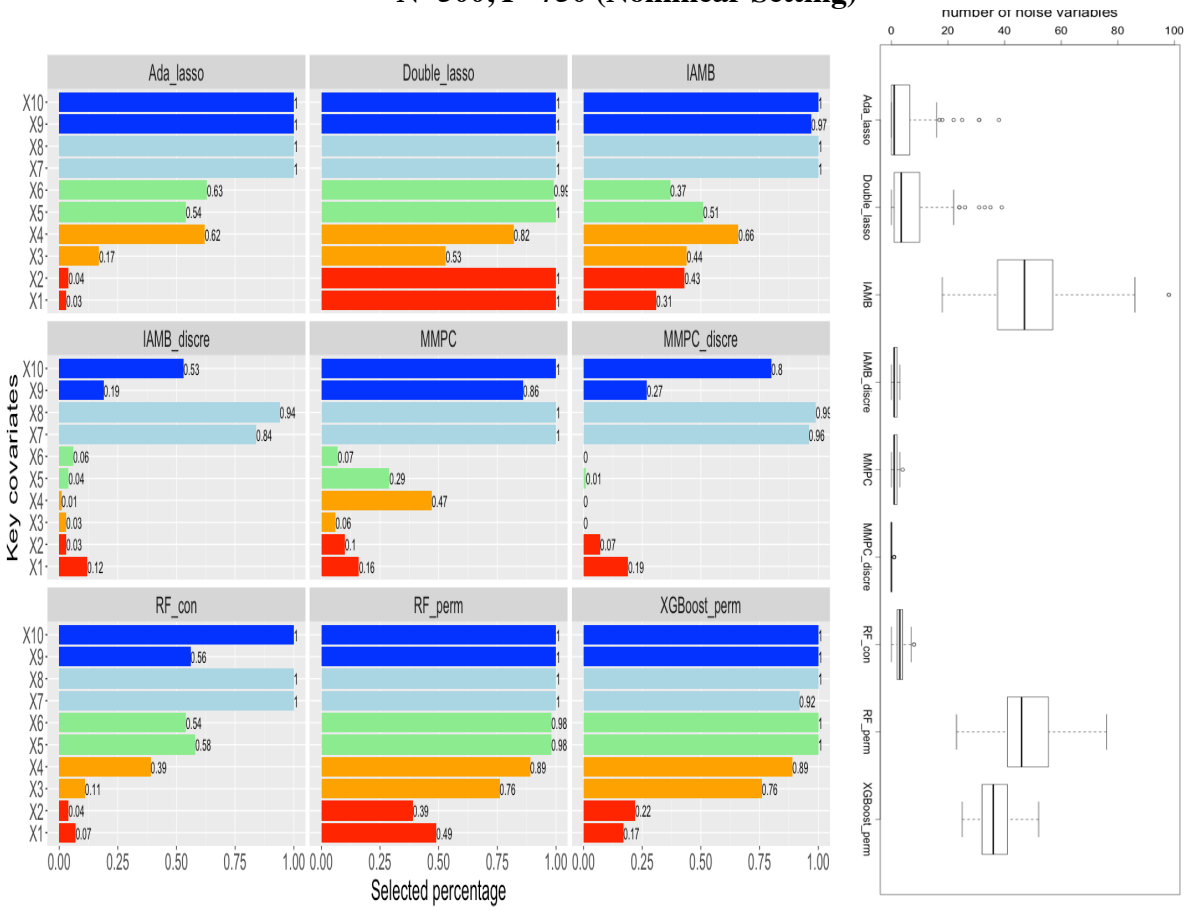


Figure 4.12: Results for the nonlinear case with a medium sample size (N = 500) and extreme high dimensionality (P = 750). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 , in red, are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

Under sample size N=500 and linear settings (Figure 4.6-Figure 4.8), Adaptive-Lasso performed the best with an average rate of 0.99 for including the key covariates and an average rate of 0.05for including the instrumental variables. IAMB also performed well with an average rate of 1.00including the target covariates and a slightly higher average rate of 0.10 including the instrumental variables. IAMB also included a higher proportion of noise variables (median = 48) under the high dimension setting (P = 750). Here, Double-lasso included almost 100% of the instrumental variables. Weak confounding variable were difficult for tree-based methods to detect (consult Figure 4.6-Figure 4.8). Among tree-based methods, permutation based random forest was best when P=34 and permutation-based regularized XGBoost's performance was best when dimensionality was higher.

Under sample size N=500 and nonlinear setting (Figure 4.9- Figure 4.12), Tree based ensemble methods performed best among all methods. Permutation-based random forest had an average rate of 0.95 including all key covariates and average rate of 0.29 including instrumental variables. Permutation-based XGBoost had an average inclusion rate of around 0.90 for key covariates and 0.12 for instrumental variables. When dimensionality was low, permutation-based random forest had the best average rate, which is 0.94, for including the target variables. When P=250 and P=750, permutation-based regularized XGBoost was best with an average rate of 0.88 for including the key covariates. Bayesian network-based methods tend to include more instrumental variables and were unable to detect variables involved in interaction terms; regularized regression methods were also less likely to include interaction terms. When P=750, IAMB and permutation based random forest included the most noise variables among all approaches

4.2.3 Large Sample Size

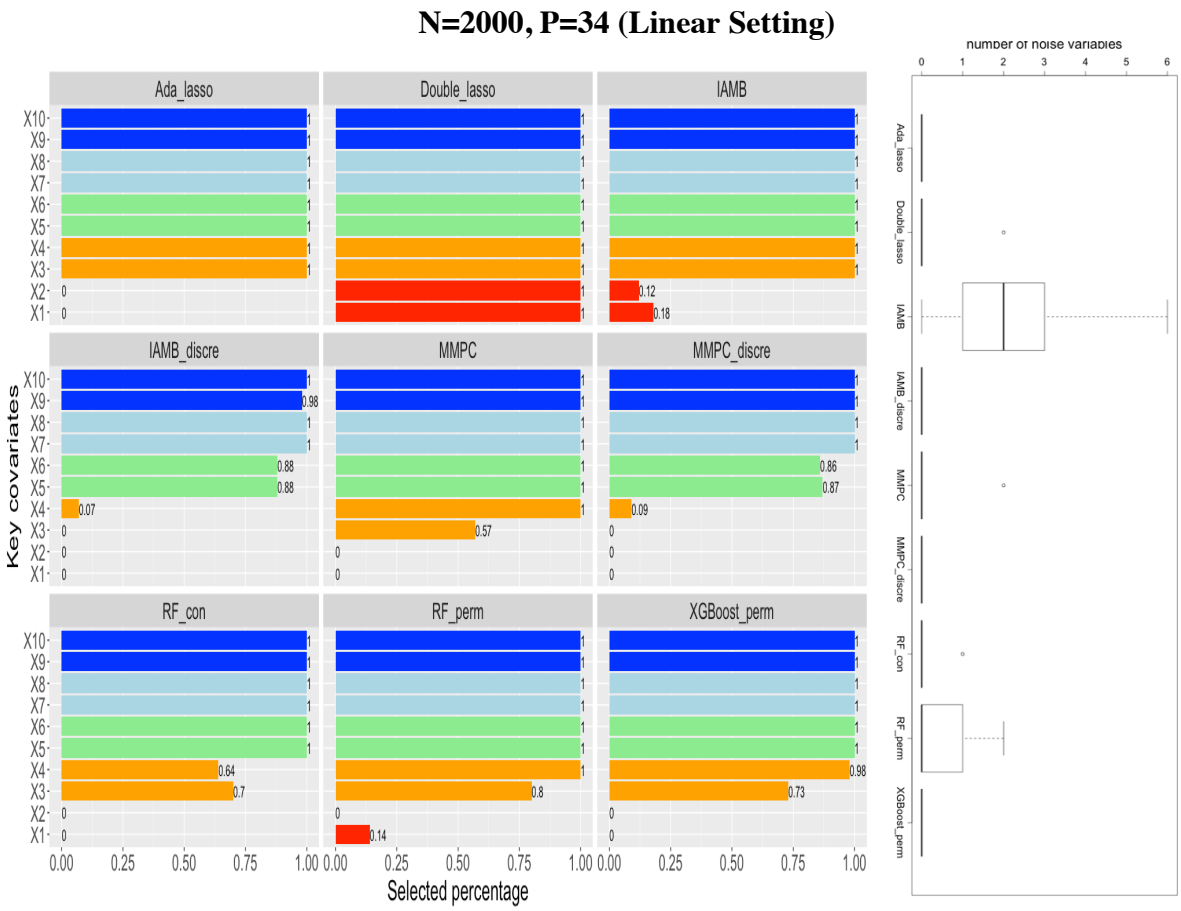


Figure 4.13: Results for the linear case with a large sample size (N = 2000) and low dimensionality (P = 34). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 , in red, are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

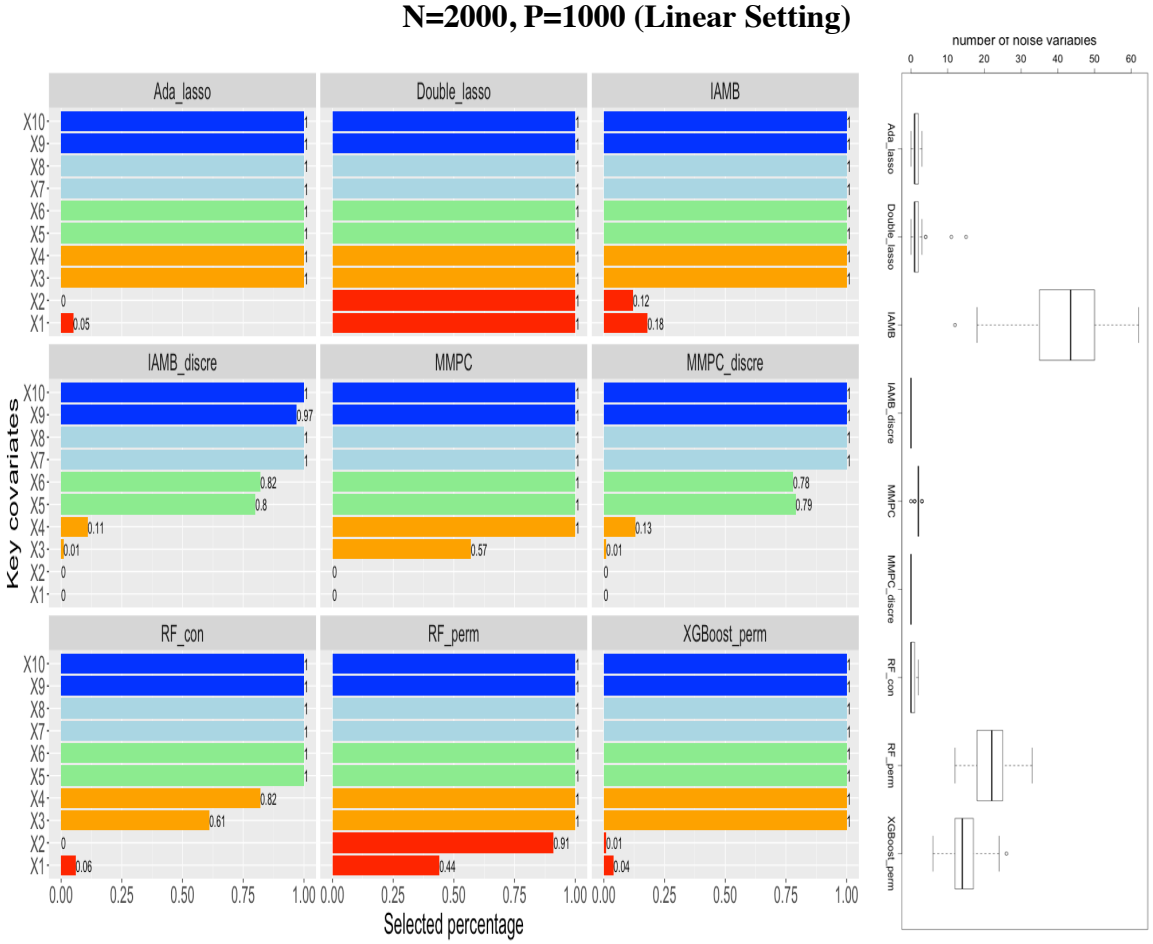


Figure 4.14: Results for the linear case with a large sample size (N = 2000) and high dimensionality (P = 1000). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2, in red, are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

N=2000, P=3000 (Linear Setting)

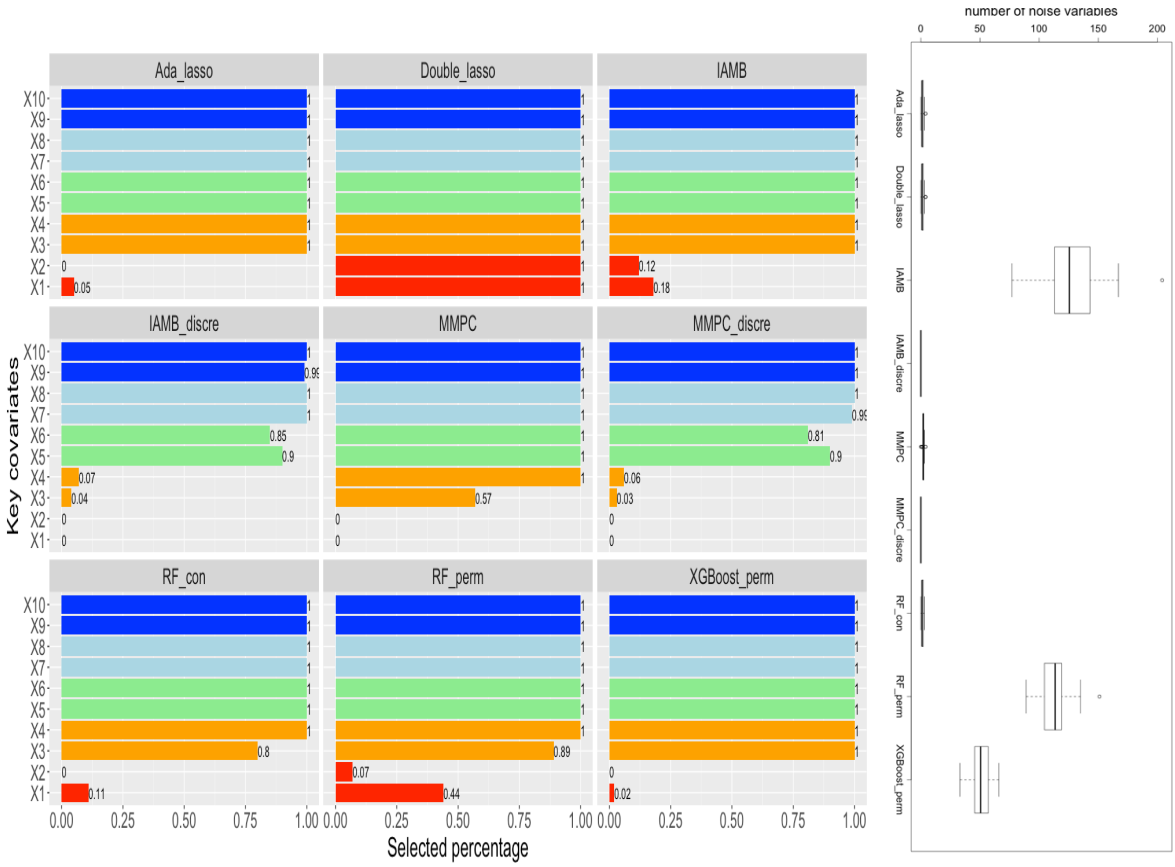


Figure 4.15: Results for the linear case with a large sample size (N = 2000) and extreme high dimensionality (P = 1000). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 , in red, are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

N=2000, P=34 (Nonlinear Setting)

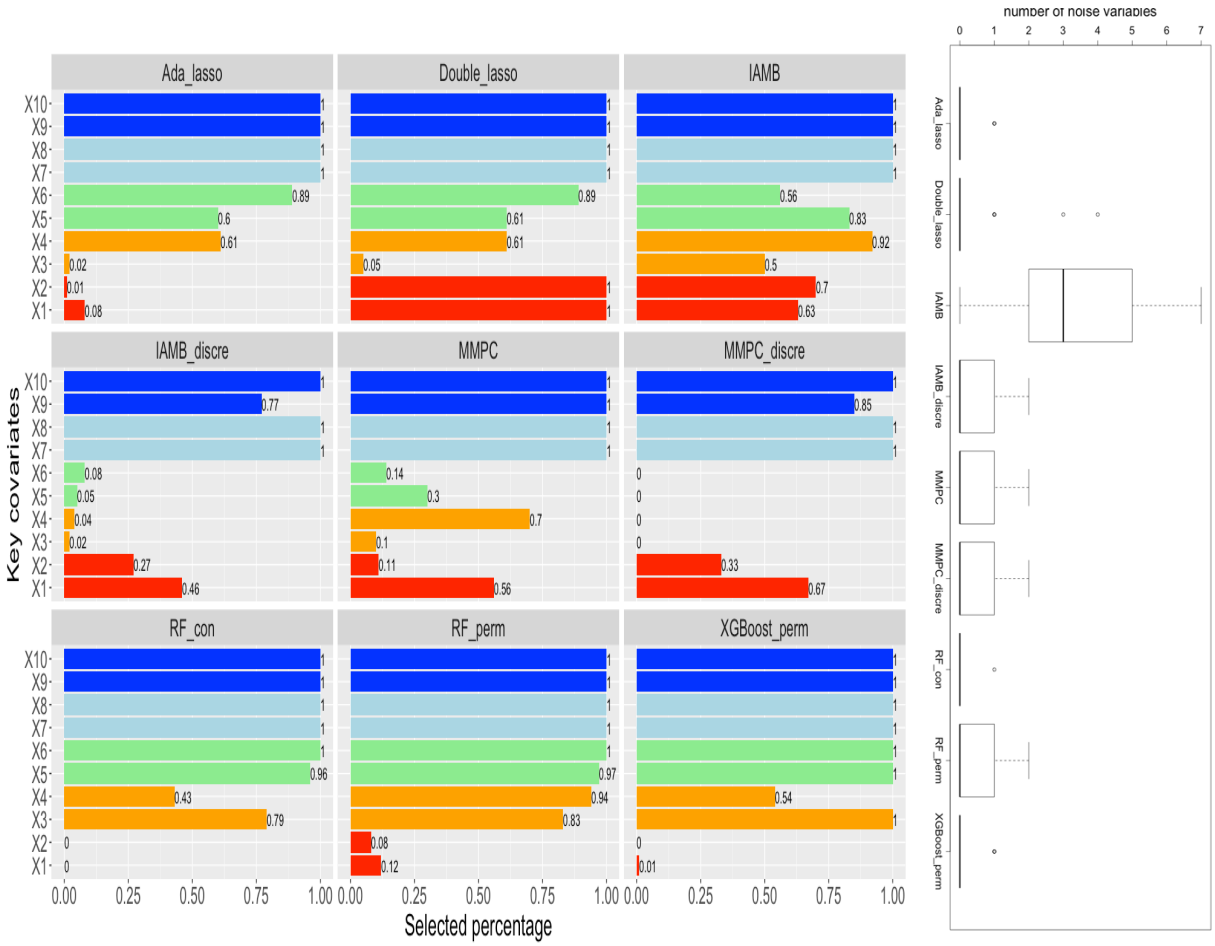


Figure 4.16: Results for the nonlinear case with a large sample size (N = 2000) and low dimensionality (P = 34). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2, in red are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

N=2000, P=1000 (Nonlinear Setting)

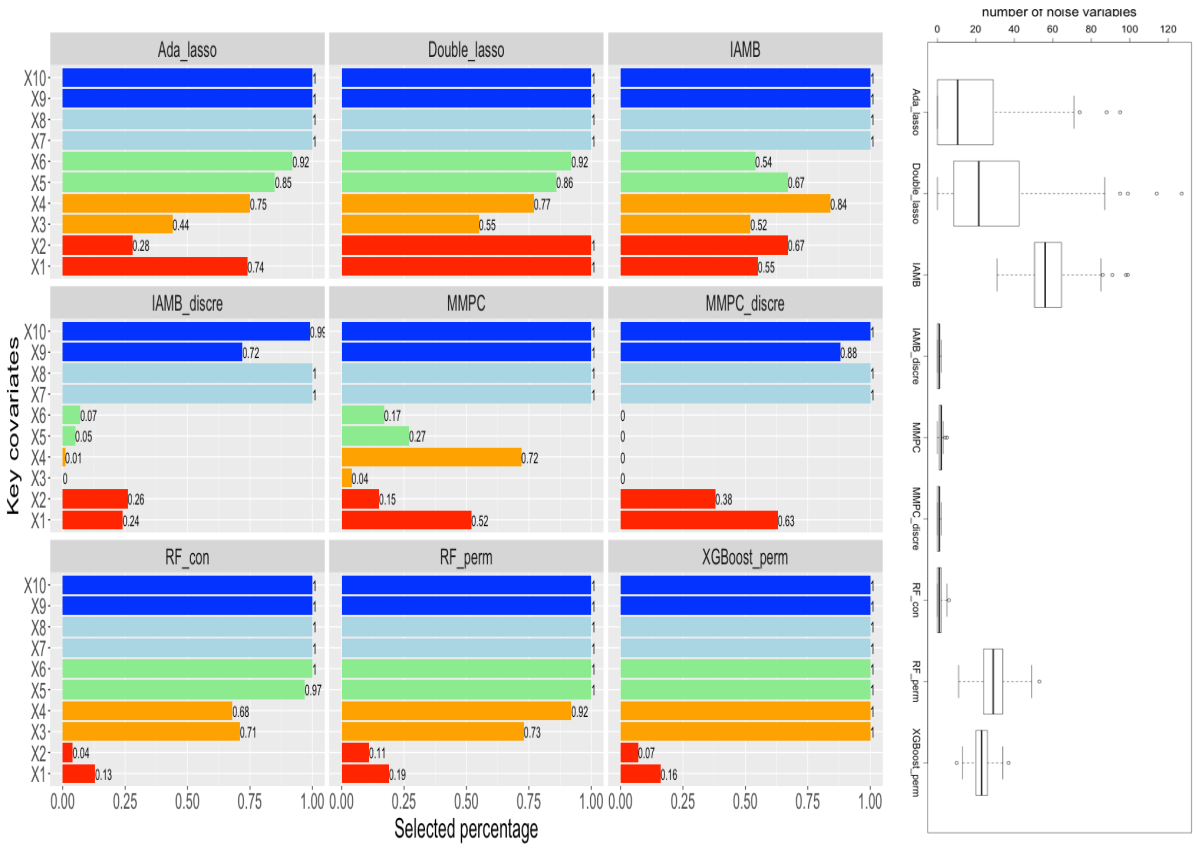


Figure 4.17: Results for the nonlinear case with a large sample size (N = 2000) and high dimensionality (P =1000). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2, in red, are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

N=2000, P=3000 (Nonlinear Setting)

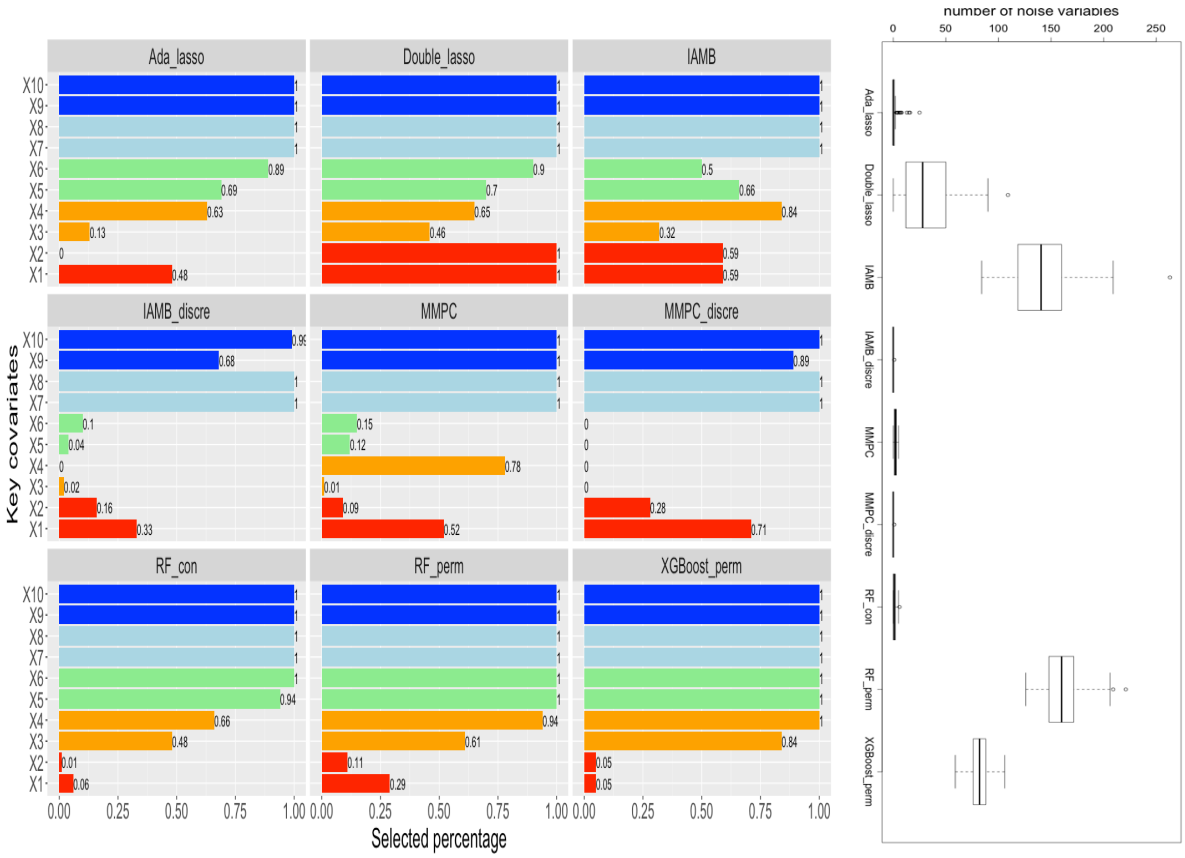


Figure 4.18: Results for the nonlinear case with a large sample size (N = 2000) and extreme high dimensionality (P =3000). Selected percentage is the percent of replications (out of 100) for which the target variable was selected. Left panel -- For each target variable in X_1 through X_{10}, the bar charts display the percentage of simulation replications (out of 100) for which the target variable was selected. X_1 and X_2 , in red, are instrumental variables and should only be selected by the double lasso method; all remaining target variables (i.e., X_3 through X_{10}) should be selected by all methods. Variables X_3 and X_4, in yellow, are weak confounding variables; X_5 and X_6, in green, are moderately strong confounding variables; X_7 and X_8, in light blue, are strong confounding variables; and X_9 and X_{10}, in blue, are risk factors. Right panel -- Boxplots display the distribution of the number of noise variables selected across 100 simulation replications. Noise variables are defined as all variables excluding $X_1 \ldots X_{10}$; that is, noise variables were not involved in the data-generating process.

Under N=2000 and a linear setting (Figure 4.13-Figure 4.15), the difference between methods with respect to detecting key covariates is minimal. All methods except the discretized Bayesian networks had above a 0.94 average rate of detecting key covariates. The penalized regression approaches (i.e., permutation-based regularized XGBoost and Double Lasso) and IAMB had an average rate of 1.00 for identifying key covariates. The major differences across methods at this largest sample size setting were with respect to the proportion of noise and instrumental variables incorrectly retained. Recall that all the methods except double lasso are intended to ignore (i.e., not select) instrumental variables. In this largest sample size setting, IAMB worked well with low dimensionality (i.e., P=34). When dimensionality increased, more noise variables were selected by IAMB. Tree-based methods also selected a high proportion of noise variables.

Permutation-based regularized XGBoost had an average rate of 0.46 and conditional random forest had an average rate of 0.35 for including instrumental variables. Both are lower than the permutation based random forest which had an average rate of 0.61 for including instrumental variables at this highest sample size. permutation-based regularized XGBoost and conditional random forests performed better than traditional permutation-based random forests for the extremely high dimensionality case (P=3000); traditional permutation-based random forests included a median value of 120 noise variables in the final covariate sets for this scenario, whereas permutation-based regularized XGBoost and conditional random forests included a median value of 48 and 10.

Under N=2000 and nonlinear setting (Figure 4.16-Figure 4.18), only ensemble tree-based methods were able to identify interaction terms. Permutation-based random forests and permutation-based XGBoost had the best inclusion rates for key covariates, with average inclusion rates of 0.96 and 0.97, respectively. Conditional random forests had an average inclusion rate of

0.90 for key covariates. For regularized regression-based methods, the average inclusion rates for key covariates were around 0.82. For Bayesian net-based methods, the average inclusion rates were lower than 0.50. Permutation-based XGBoost and conditional random forests included the smallest proportions of noise and instrumental variables. Permutation-based regularized XGBoost had an average rate of 0.06 and conditional random forest had an average rate of 0.04 for including instrumental variables. The median noise number for Permutation-based regularized XGBoost included a median value of 52 noise variables and the median noise number for conditional random forests is 15. When dimensionality was low (P=34), permutation-based random forest performed the best. When dimensionality was higher than 34, permutation-based XGBoost and conditional random forests were better alternatives. Discretized Bayesian nets could not identify interaction terms with a sample size of N=2000 in this simulation setting: they had an average rate of 0.58, including key covariates.

4.3 Result for Study III

4.3.1 Empirical Monte Carlo Simulation with 34 Pre-identified Key Covariates

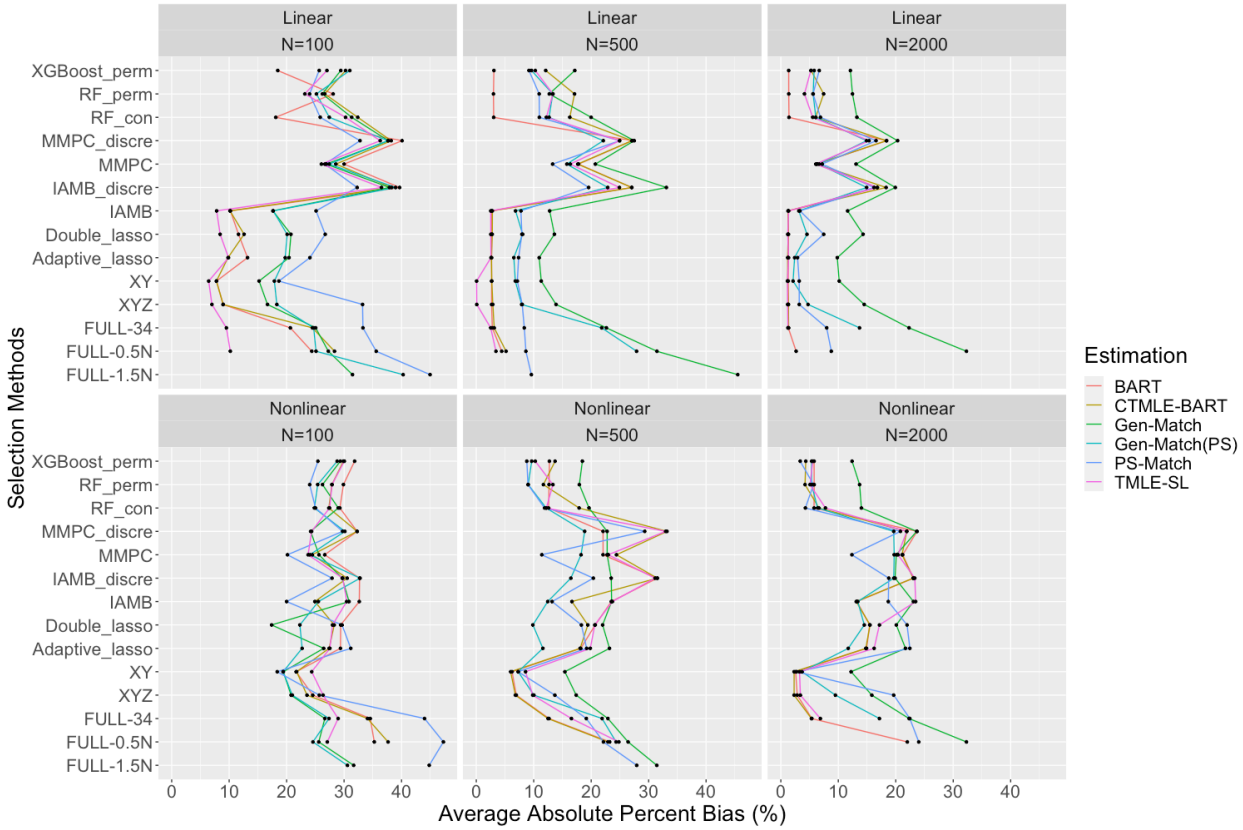


Figure 4.19: Displays the average absolute percent bias over 100 replications of covariate selection methods with different estimation strategies. For covariate selection methods, XY and XYZ are oracle sets based on the outcome and disjunctive cause criteria, respectively; FULL-34 is all 34 covariates; FULL-0.5N include 0.5N covariates which contains 34 key covairates and other noise covairates; FULL-1.5N include 1.5N covariates which contains 34 key covairates and other noise covairates

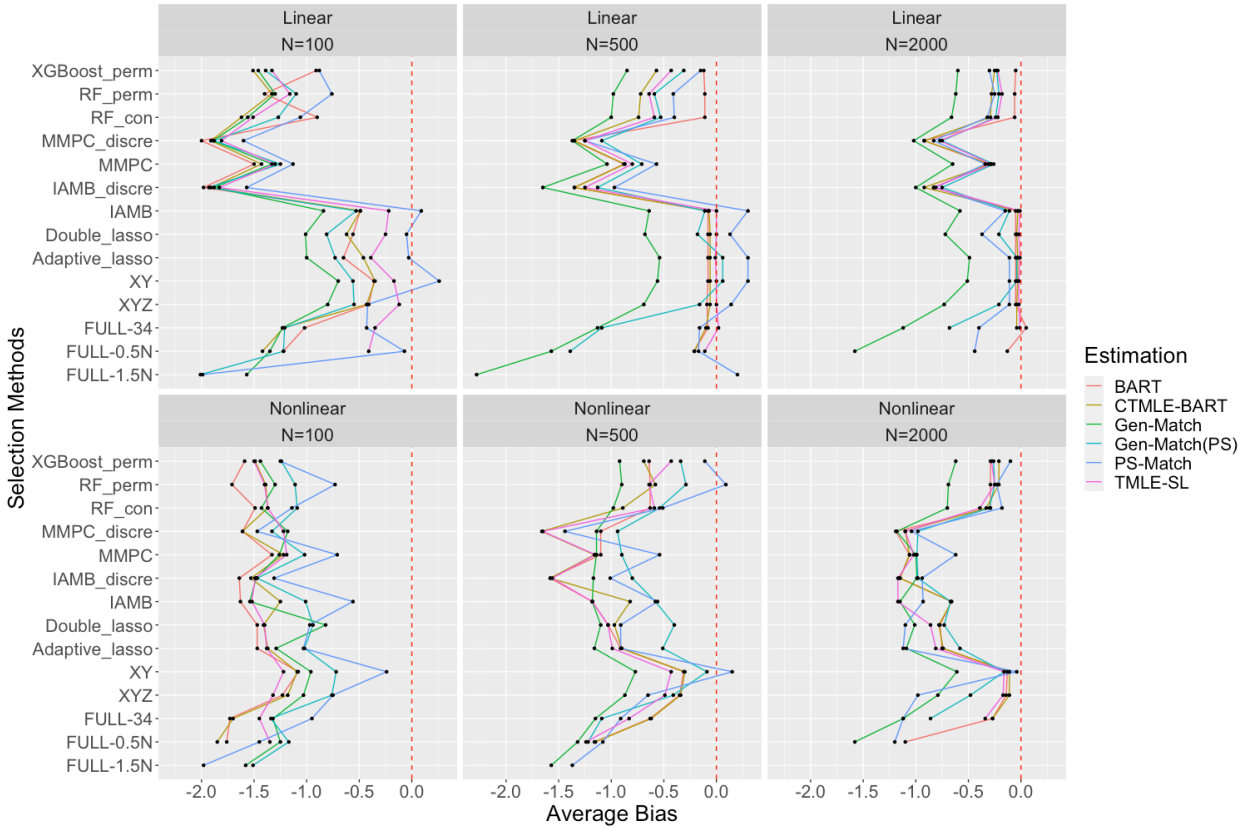


Figure 4.20: Displays the average bias (over 100 replications) of different covariate selection methods with different estimation strategies. For covariate selection methods, XY and XYZ are oracle sets based on the outcome and disjunctive cause criteria, respectively; FULL-34 is all 34 covariates; FULL-0.5N include 0.5N covariates which contains 34 key covairates and other noise covairates; FULL-1.5N include 1.5N covariates which contains 34 key covairates and other noise covairates

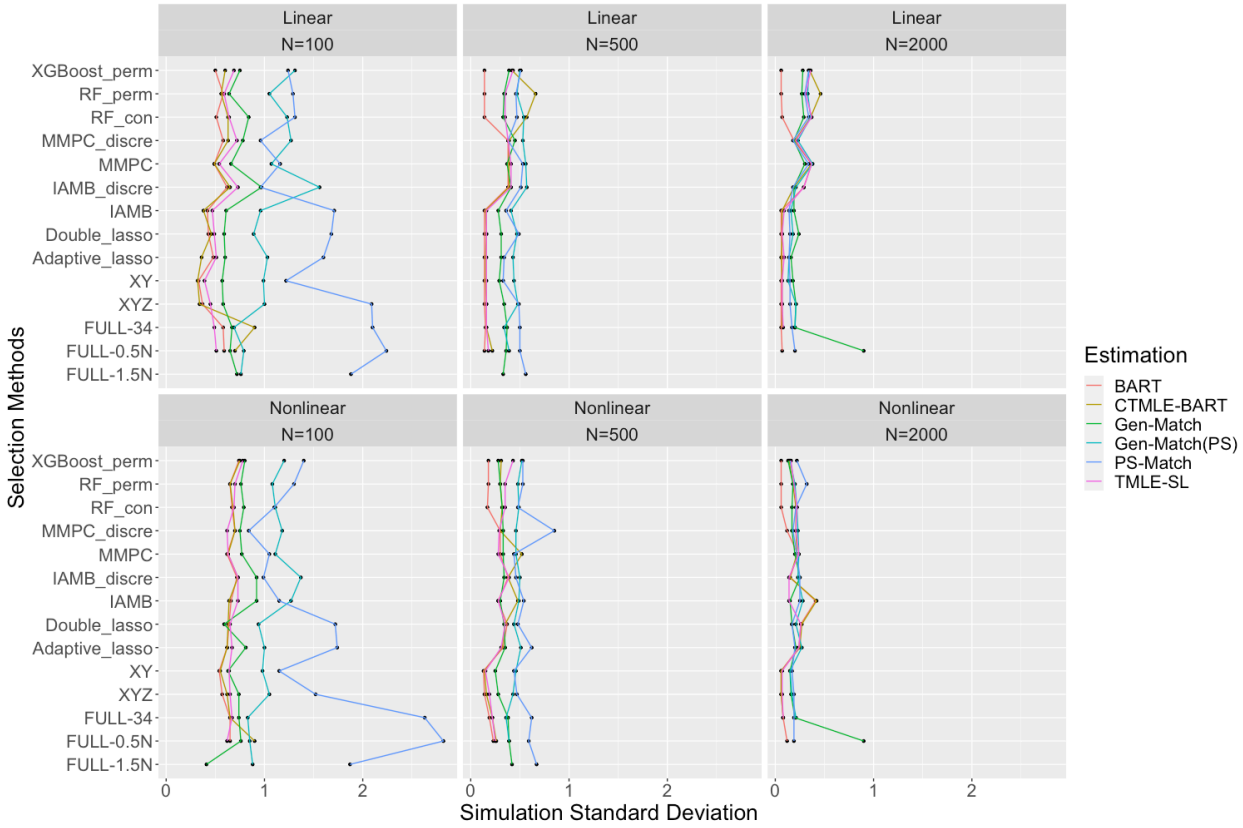


Figure 4.21: Displays the simulation standard error (over 100 replications) of different covariate selection methods with different estimation strategies. For covariate selection methods, XY and XYZ are oracle sets based on the outcome and disjunctive cause criteria, respectively; FULL-34 is all 34 covariates; FULL-0.5N include 0.5N covariates which contains 34 key covairates and other noise covairates; FULL-1.5N include 1.5N covariates which contains 34 key covairates and other noise covairates.

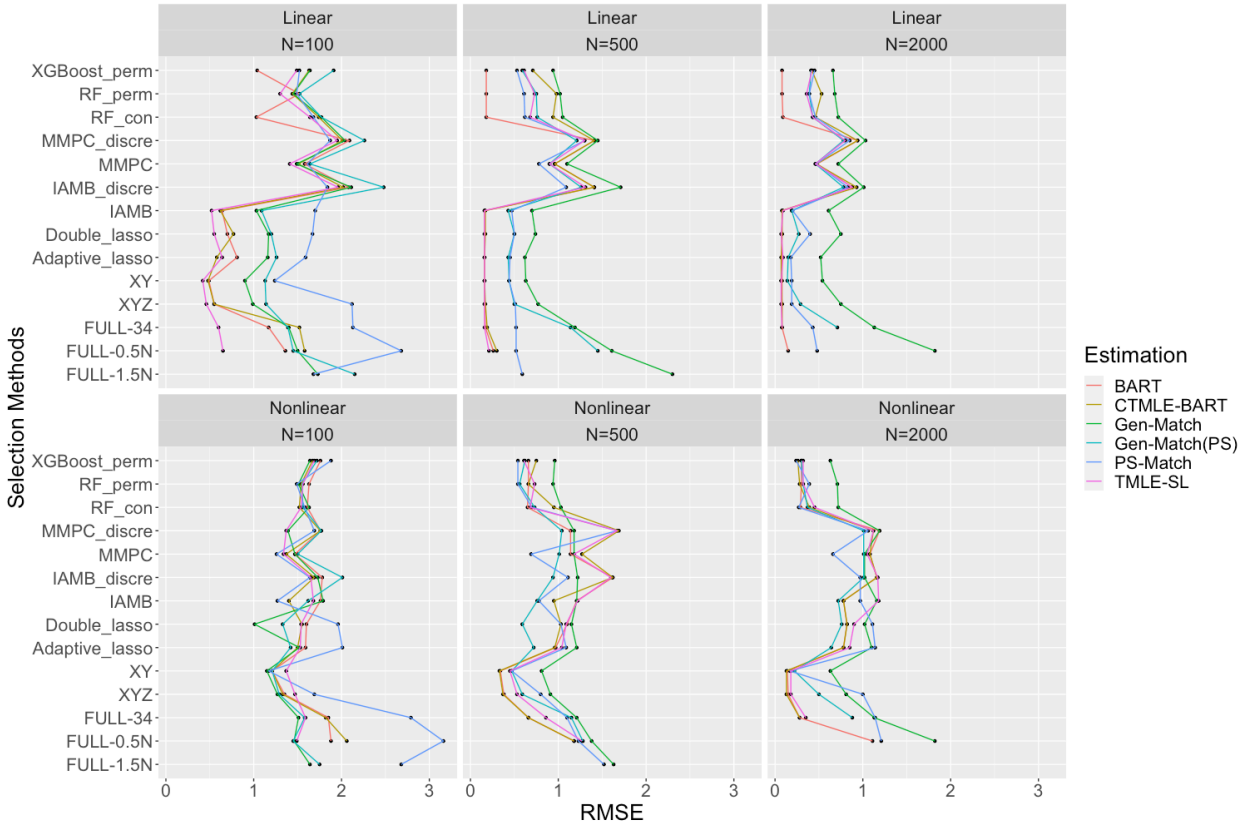


Figure 4.22: Displays the root mean square error (RMSE over 100 replications) of different covariate selection methods with different estimation strategies. For covariate selection methods, XY and XYZ are oracle sets based on the outcome and disjunctive cause criteria, respectively; FULL-34 is all 34 covariates; FULL-0.5N include 0.5N covariates which contains 34 key covairates and other noise covairates; FULL-1.5N include 1.5N covariates which contains 34 key covairates and other noise covairates

Recall that in Study I, all estimation methods were run with full covariate sets, to the extent possible given computational limitations. Also recall that in Study II, all covariate selection methods were run on all simulation cells. Ideally, in Study III, all estimation methods examined in Study I would be crossed with all selected variable sets returned from Study II to measure performance across all cells of the simulation. However, due to the extreme computational cost associated with running the estimation strategies with high dimensional data, covariate selection

output and estimation strategies were only crossed and run together for the case where P = 34 (i.e., 10 primary variables plus 24 noise variables) in Study III.

To interpret the output shown in Figures 4.19-4.22, consider the bottom-middle plot in Figure 4.22, which represents the RMSE for the nonlinear case when sample size was N = 500. For that plot, the minimum RMSE across all methods and covariate selection schemes was achieved at a value of 0.34 by BART and 0.33 for CTMLE (they are overlapping in the figure) for the XY oracle set. Since the oracle set is not known in practice, the most relevant question is how estimation with the full set of 34 covariates (FULL-34) performs relative to estimation after data-based covariate selection.

When using all 34 covariates (as was done in Study I), the RMSE under BART was 0.66, CTMLE was 0.66, TMLE-SL was 0.86, PS-Match was 1.10, Genetic-Match was 1.21 and Genetic-Match-PS 1.15. When coupled with variable selection under the tree-based methods, for example, RMSE for BART after XGBoost_perm, RF_perm, and RF_con was 0.66, 0.66, and 0.65, respectively and RMSE for CTMLE after XGBoost_perm, RF_perm, and RF_con was 0.75, 0.66, 0.95, respectively.

Although the average causal effect was not estimated with covariate sets selected for higher dimensional data, the P = 34 case may be used as a best-case scenario for comparison. Admittedly, covariate sets selected with higher dimensionsal data will contain more noise variables, but as was observed in Study II, this depends greatly on the method and, in any case, the number of noise variables retained. For instance, even in the highest dimension and largest size cases, less than one third of one percent of noise was retained.

To make meaningful comparisons for the higher-dimensional cases (P = 0.5N and P = 1.5N) across sample sizes, with the caveats discussed in the previous paragraph notwithstanding, baseline

results from Study I are presented as well for the higher dimensional cases. Continuing to use the bottom-middle plot in Figure 4.22 as an example, note that the FULL-0.5N condition represents the RMSE when each estimation method was run with all 500*0.5 = 250 covariates, as was done in Study I. Consulting Table 4.2, one can see that for N = 500 and P = 0.5N, the RMSEs for the estimation methods were as follows: BART = 1.18, CTMLE = 1.18, TMLE-SL = 1.27, PS-Match = 1.23, Genetic-Match = 1.38 and Genetic-Match-PS=1.28. Thus, comparing these values to the best RMSE values after covariate selection in this scenario, it becomes clear that there is potential for meaningful improvments due to covariate selection.

Under linear settings, IAMB, Double Lasso and Adaptive Lasso performed best in terms of reducing bias across all sample sizes. Under moderate sample sizes (N=500, N=2000) and linear settings, tree-based covariate selection methods also peformed well. Under a small sample (N=100) and nonlinear setting, there is not much difference across covariate selection methods in terms of reducing bias, except double lasso, which had the highest bias. Under the moderate sample size conditions (N=500, N=2000) and nonlinear settings, tree-based approaches outperform all others. Discretized Bayesian nets (i.e., MMPC-discre and IAMB-discre) resulted in the largest biases. The combination of approporate covariate selection methods with BART, CTMLE-BART and TMLE-SL were best in terms of bias and RMSE in moderate sample size cases with high dimensionality, although note that in some cases, there were no comparators for bias and RMSE because estimation with the full covariate set was not possible. With low dimesionality (P=34) and large sample size (N = 2000), the usefulness of covariate selection in terms of bias controlling is not obvious for certain methods (i.e., BART, CTMLE-BART, TMLE-SL).

From Figure 4.21, the matching propensity scores had the highest standard errors in comparison to other approaches under both linear and nonlinear settings. The appropriate covariate selection

could help reduce the simulation standard error propensity score matching estimates especially when N=100. For BART, CTMLE-BART, TMLE-SL, the role of covariate selection in terms of reducing standard error was not obvious. These estimation strategies had a relatively smaller standard error even without covariate selection.

From Figure 4.22, pre-processing was essential to maintain small RMSE when a sample size was small and/or when dimensionality is high. Under a linear setting, Lasso based, and Bayesian networks could successfully reduce RMSE. When a sample size grows (N>500), the tree-based methods also performed well. Under a nonlinear setting, tree-based methods were the best. The combination of these covariate selection methods with BART based estimation strategies (i.e., BART, TMLE-SL and CTMLE-BART) could help maintain a low RMSE.

4.3.2 Empirical Study with 34 Pre-identified Covariates

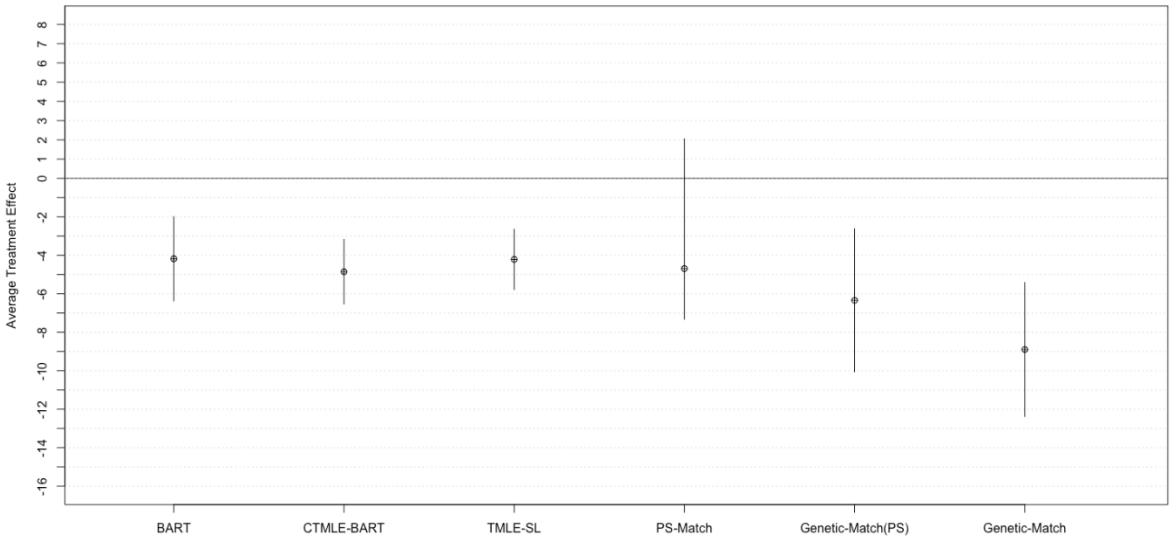


Figure 4.24: 95% confidence interval for ATE with different estimation strategies.

Methods		**All Covariates**
BART	Point Estimate	-4.18
	95% C. I	[-6.37, -1.99]
TMLE-SL	Point Estimate	-4.85
	95% C. I	[-6.53, -3.17]
CTMLE-BART	Point Estimate	-4.21
	95% C. I	[-5.78, -2.64]
PS-Match	Point Estimate	-4.69
	95% C. I	[-7.32, 2.06]
Genetic-Match (PS)	Point Estimate	-6.34
	95% C. I	[-10.06, -2.62]
Genetic-Match	Point Estimate	-8.90
	95% C. I	[-12.38, -5.41]

Table 4.3: ATE estimated with 34 covariates by different estimation methods.

From Table 4.3, the point estimation of ATE from BART, TMLE-SL, CTMLE-BART and PS-Match were very close. Genetic-Match and Genetic-Match (PS)'s point estimation deviated with other approaches. All point estimates indicated there was a negative average treatment effects of special education on mathematic performances. From Figure 4.23, the confidence interval based on BART, TMLE-SL, CTMLE-BART are narrower relative to matching and there was a slight difference across matching methods. First, Genetic matching produced a slight narrow confidence interval in comparison to PS-Match. Second, PS-Match's confidence interval contained zero, which indicated that no statistically significant conclusion could be drawn from the point estimates. However, Genetic-Match's confidence interval contained non-zero and was located at negative sides, which resulted in a similar conclusion as other machine learning based estimation approaches.

Methods		FULL	IAMB	MMPC	IAMB-discr	MMPC-discr	RF-perm	RF-con	XGBoost-perm	Ada-lasso	Double-lasso	Consensus
BART	Point Estimates	-4.18	-4.33	-5.00	-5.73	-5.83	-2.45	-4.92	-5.22	-5.03	-4.23	-4.36
	95% C.I	[-6.37, -1.99]	[-6.56, -2.11]	[-7.03, -2.97]	[-8.00, -3.45]	[-7.06, -3.97]	[-4.52, -0.39]	[-7.27, -2.57]	[-7.45, -3.00]	[-7.42, -2.65]	[-6.93, -1.54]	[-6.46, -2.26]
TMLE-SL	Point Estimates	-4.85	-4.73	-5.45	-6.07	-6.07	-2.61	-4.74	-5.61	-5.18	-4.85	-4.26
	95% C.I	[-6.53, -3.17]	[-6.31, -3.16]	[-7.24, -3.67]	[-7.80, -4.34]	[-7.81, -4.34]	[-4.35, -0.88]	[-6.41, -3.92]	[-7.31, -3.92]	[-6.91, -3.44]	[-6.54, -3.15]	[-5.05, -3.47]
CTMLE-BART	Point Estimates	-4.21	-4.58	-4.2	-5.52	-5.11	-2.38	-5.09	-5.32	-5.49	-4.58	-4.46
	95% C.I	[-5.78, -2.64]	[-6.77, -2.98]	[-5.76, -2.64]	[-7.06, -3.97]	[-6.06, -3.61]	[-4.02, -0.74]	[-6.64, -3.54]	[-7.00, -3.64]	[-7.07, -3.91]	[-6.17, -2.99]	[-6.10, -2.83]
PS-Match	Point Estimates	-4.69	-3.44	-4.63	-5.52	-5.80	-4.78	-2.73	-2.74	-5.06	-4.21	-2.91
	95% C.I	[-7.32, 2.06]	[-6.01, -0.85]	[-7.10, -2.16]	[-8.05, -2.98]	[-8.29, -3.31]	[-7.47, -2.19]	[-5.37,0.09]	[-5.28, -0.19]	[-7.65, -2.48]	[-6.80, -1.65]	[-5.48, -0.35]
Genetic-Match (PS)	Point Estimates	-6.34	-8.89	-7.47	-5.36	-13.17	-6.81	-8.16	-4.45	-9.49	-6.15	-7.40
	95% C.I	[-10.06, -2.62]	[-12.33, -5.46]	[-10.55, -4.39]	[-8.32, -2.40]	[-16.66, -9.69]	[-10.69, -2.92]	[-12.11, -4.19]	[-7.49, -1.40]	[-12.98, -6.01]	[-9.74, -2.55]	[-10.83, -3.96]
Genetic-Match	Point Estimates	-8.89	-8.49	-9.87	-7.37	-5.00	-8.56	-11.22	-8.08	-8.97	-7.67	-10.85
	95% C.I	[-12.38, -5.41]	[-11.48, -5.50]	[-13.02, -6.72]	[-10.11, -4.64]	[-7.78, -2.23]	[-11.61, -5.50]	[-15.18, -7.26]	[-11.60, -5.10]	[-12.24, -5.71]	[-10.97, -4.36]	[-14.00, -7.71]

Table 4.4: Displays a 95% confidence interval for ATE with different covariate selection methods crossing over different estimation strategies. Consensus included covariates that were least selected by five out of nine covariate selection methods (i.e., consensus selection).

By assuming ignorability was satisfied by including all 34 covariates (i.e., FULL), a successful feature selection procedure should (1): produce similar point estimates (2): result in narrower confidence interval (i.e., smaller standard error). From table 4.4, the machine learning based estimation approaches were not very sensitive to covariate selection in comparison to matching based approaches. The selection methods that significantly change the point estimates and confidence intervals were discretized Bayesian network (i.e., MMPC-discre and IAMB-discre) and permutation based random forests. In contrast, matching approaches were sensitive to various covariates selection methods especially for Genetic-match. Covariate selection successfully narrowed down the confidence interval for propensity score matching and resulted in a similar conclusion as machine learning based estimation approaches. Pre-processing by using discretized Bayesian networks and certain tree-based approaches (i.e, RF-con and XGB-perm) created different point estimates and confidence interval in comparison to using other pre-processing methods. In general, consensus methods produced the most stable estimates and confidence intervals across different estimation methods.

Covariates Name	Ada_lasso	Double_lasso	XGBoost_perm	RF_perm	RF_con	MMPC	MMPC_discre	IAMB	IAMB_discre
DEMOGRAPHIC									
GENDER	X	X	X	X	X	X	X	X	X
WKWHITE	X	X				X	X	X	
WKSESL	X	X	X	X	X	X	X	X	X
ACADEMIC									
RIRT		X		X			X		X
MIRT	X	X	X	X	X	X	X	X	X
S2KPUPRI	X	X	X		X			X	X
P1EXPECT								X	
P1FIRKDG	X	X	X	X	X			X	
P1AGEENT		X	X	X	X		X	X	
ApprchT1	X	X				X		X	X
P1HSEVER	X	X	X			X		X	
SCHOOL COMPOSITION									
avg_MIRT								X	
avg_SES	X	X				X		X	
avg_ApprchT1								X	
S2KMINOR		X						X	
FAMILY CONTEXT									
P1FSTAMP	X	X		X		X	X	X	
ONEPARENT	X	X				X		X	X
TEPPARENT		X							
P1HMAFB						X		X	
HEALTH									
P1EARLY						X		X	
P1WEIGHO								X	
C1FMOTOR	X	X	X	X	X	X		X	
P1SADLON		X							
P1IMPULS		X							
P1ATTEN		X							
P1SOLVE	X	X		X		X			
P1DISABL		X							
P1PRONOU	X	X			X			X	
TOTAL	14	22	8	9	8	13	7	21	7

Table 4.5: Covariates being selected in empirical study.

Table 4.5 displays covariates selected by different covariate selection methods. The dimensionality of the covariate sets after covariate selection ranged from 7 to 22, with a median value of 9. A tree-based and discretized Bayesian-net based method retained the smallest number of covariates. Double lasso retained the largest number of covariates. The covariates were marked such that it is

at least selected by five out of nine covariate selection methods (i.e., consensus selection). These covariates were: Gender, Race, Social Economic Status (SES), Kindergarten Math Score, Public School, First-Time Kindergartener, Child's Age at K Entry (Months), Approaches to Learning Rating, Attended Head Start, Received Food Stamps, One-Parent Family and Fine Motor Skills. These covariates should receive more attention than others since they represented the more generalized agreement among different approaches and produced more stable estimates (see Table 4.4).

The covariates that were selected by other approaches, but not by tree-based approaches were: P1HSEVER, ApprchT1, WKWHITE and Avg-SES. That may be due to the fact that there were interaction terms that were not detected by methods which assumed linearity. RIRT was only eliminated by methods using conditional independence or its equivalence. For example, RIRT was selected by RF-perm but not by RF-con, XGBoost-perm, IAMB and Ada-Lasso. The person's correlation between MIRT and RIRT is 0.71. After fitting a simple linear regression of MIRT and RIRT on fifth grade math scores, it was discovered that RIRT is not statistically significant. This finding indicated that RIRT was marginally related to math score. Subsequently, this shows why covariate selection methods that aimed to identify strong relevant features eliminate RIRT.

Chapter 5: Summary and Discussion

The purpose of this study was to explore the role of feature selection in high dimensional causal inference. To fulfill this purpose, the previous chapters covered three simulation studies and one empirical study to resolve the existing problems. First, there was a discussion on the main findings from the simulation and empirical studies. The discussion followed the order of questions raised in Chapter 2. Then, there was a discussion on the empirical implications of the findings toward future educational research. Finally, there was a description of a possible extension of this study and clarification of the limitations.

5.1 Discussion

5.1.1 Study I: Is covariate selection necessary when feature selection is embedded in causal estimation?

To answer this question, researchers need to know: (1) How large is the sample size? (2) What is the dimensionality of the covariate sets? (3) How complex is the relationship among covariates, exposure variable and outcome variable? (4) Which estimation strategy is used for causal estimation? Recall that Study I focused on a case where estimation strategies were used with the full set of covariates; that is, no prescreening via filter covariate selection was done.

With a relatively large sample size (N=2000) and small dimensionality (P=34), BART, CTMLE-BART and TMLE-SL (i.e., methods model response surface or double robust) could achieve estimates close to the oracle estimation without covariate selection in most conditions. Under particular circumstances (i.e., very complex functional form of response surface and assignment mechanism), covariate selection methods may help the estimation of BART, CTMLE-BART and TMLE-SL close to the oracle estimation (see, for example, bottom left of Table 4.1).

Covariate selection was more important for genetic matching (non-parametric matching) and propensity score matching (i.e., methods that model the assignment mechanism).

With a relatively large sample (N=2000), when the dimensionality of the covariate increased (i.e., P=0.5N and P=1.5N), covariate selection became necessary for all estimation approaches. Being that some methods have a very long computing time (CTMLE-BART, TMLE-SL), some methods (BART, PS-Match, Genetic-Match and Genetic-Match(PS)) had a relatively large bias, and subsequently some methods could not operate under P >N case (BART, PS-Match ,Genetic -Match and Genetic-Match(PS)).

With a relatively small sample size (N=500, N=100), the role of covariate selection was essential but varies under different circumstances. When a response surface/assignment mechanism is complex and a sample size was extremely small (i.e., N=100), covariate selection methods may not be helpful for the estimation strategies explored in this dissertation. Nonetheless, with a simpler response, the surface/assignment mechanism and/or an increase of the sample size (N=500), choosing the appropriate covariate selection methods could greatly reduce the estimation bias and reduce the simulation standard error (see, for example, right middle cell of Table 4.1).

Regarding the complex relationship of the assignment mechanism/response surface, most estimation strategies had a relatively large bias and high simulation standard error when the sample size was N = 100. Under this circumstance (i.e., small sample), estimation approaches performed differently. When a response surface/assignment mechanism was linear, TMLE-SL had superior performances, even when the sample size was extremely small (N=100). However, when the response to a surface/assignment mechanism became more complex, TMLE-SL was not as effective as using BART alone. The reason may come from the composition of the super learning library, including methods (e.g., Lasso) that had very good performance under linear settings but

did not perform well under nonlinear settings. Therefore, choosing the correct composition of a super learning library was essential. CTMLE-BART outperformed BART only when dimensionality was relatively high (P=0.5). However, CTMLE-BART could not handle the extreme case (P=1.5N). Genetic matching could outperform another estimation approach under a very small sample (N=100) and complex relationship case.

Perhaps the most important factor to note here was that there are many empty cells in the simulation results. These NA cells represented cases where estimation with the full covariate set was simply infeasible due to the computational cost associated with a particular method involving that combination of sample size and dimensionality. In these cases, which made up a majority of cells in the highest dimensional cases, estimation was not even possible without some sort of pre-screening via filter methods for covariate selection

5.1.2 Study II: How does the accuracy of covariate selection methods vary?

Under a small sample (N=100) and linear settings, IAMB, Adaptive Lasso and Double Lasso could identify key covariates successfully. However, Double Lasso would include more instrumental variables, which may have a large impact for certain causal estimations strategies. For example, with a small sample, including instrumental variables for propensity score matching may lead to pool overlap between control and treatment groups that may bias the causal estimates. Increasing the dimensionality of covariate space would not result in large problems for IAMB and Adaptive Lasso.

Under a small sample (N=100) and nonlinear setting, it's difficult for all covariate selection methods to identify key covariates. The methods with the best performance were Adaptive Lasso, Double Lasso and permutation-based random forest. Increased dimensionality would increase the

rate of the instrumental variable being selected. To include all key covariates, researchers may consider combining the covariates selected by different covariate selections methods.

Under a moderate sample (N=500, N=2000) and linear setting, IAMB, Adaptive Lasso, Double Lasso, permutation based random forest, conditional random forest, permutation-based regularized XGBoost could identify key covariate successfully. Increased dimensionality may cause trouble for IAMB and permutation-based random forest by including more noise variables. Double Lasso should be used with caution since it would include most of the instrumental variables.

Under a moderate sample (N=500, N=2000) and nonlinear setting, tree-based methods (i.e., permutation based random forest, conditional random forest and permutation-based regularized XGBoost) were the best choices in terms of including key covariates. Permutation-based random forest may include more noise variables than a conditional random forest and permutation-based regularized XGBoost when the dimensionality is high. Discretized Bayesian net (i.e., IAMB-discre and MMPC-discre) was not very helpful in detecting interaction terms in this design. In fact, due to the information lost, these two methods have the worst performances across all scenarios.

5.1.3 Study III: Which combinations of covariate selection method with estimation method is the best?

5.1.3.1 Simulation Study

Because the primary motivation in the estimation of causal effects is to estimate average treatment effects with low bias and low variance, I begin by discussing the results of Study III, which measured these primary outcomes. Then, in order to better parse and understand the reason for performance differences observed in Study III, I look to Study II for answers.

When a covariate dimension is half of the sample size (P=0. 5N), estimation strategies may face computational difficulties (i.e., could not get the estimation results) or estimates may have a

high bias and simulation standard error. When the covariate dimension is 34, using XY (i.e., include all predictors of outcome) or XYZ (i.e., include all predictors of either outcome or treatment) covariate sets could improve estimation performances.

Under a small sample (N=100) and linear setting, Adaptive Lasso, Double Lasso and IAMB were the best choice to reduce bias and maintain a low standard error. Their combination with TMLE-SL, CTMLE-BART and BART are best. From Study II, Adaptive Lasso, Double Lasso and IAMB had the highest rate of including the target variable and a relatively small number of noise variables were included. Under a small sample (N=100) and nonlinear setting, the combination of Double Lasso and genetic matching performs relatively well. This is the most difficult scenario where the role of the covariate selection is limited. From Study II, Double Lasso had the highest average rate (approximately 0.70) including key covariates. When a sample size is small, with appropriate estimation strategies (i.e., TMLE-SL, CTMLE-BART, BART and Gen-match), including instrumental variables may not be a serious problem for causal estimation.

Under a moderate sample (N=500, N=2000) and linear setting, Adaptive Lasso, Double Lasso, IAMB, permutation based random forest, conditional random forest and permutation-based regularized XGBoost all perform well in combination with different estimation strategies, especially with BART, CTMLE-BART and TMLE-SL. From study II, all of these covariate selection methods (i.e., Adaptive Lasso, Double Lasso, IAMB, permutation-based random forest, conditional random forest and permutation-based regularized XGBoost) had at least an average rate of 0.90 rate of including key covariates. Furthermore, the median number of noise variables were less than 50. Under a moderate sample (N=500, N=2000) and nonlinear setting, tree-based covariate selection approaches (i.e., permutation-based random forest, conditional random forest and permutation-based regularized XGBoost) are best when combined with BART, CTMLE-

BART and TMLE-SL. A Discretized Bayesian network-based approach (IAMB-discre and MMPC-discre) should be avoided. From Study II, under this setting, all tree-based methods had at least average rate of 0.88 of including key covariates and the maximum median number of noise variable is 245 (RF-perm) which was relatively small in comparison to the sample size. In contrast, Discretized Bayesian network-based approaches had a relatively average inclusion rate for key covariates (approximately 0.59). When a sample size is larger (N=2000) and dimensionality is low (P=34), covariate selection is not necessary for BART, CTMLE-BART and TMLE-SL.

5.1.3.1 Empirical Study

From table 4.5, the covariate selection can reduce the dimensionality of the covariate sets from 34 to as low as 7. That may indicate the key covariate that is necessary is smaller than 34. The covariates selected by five out of nine covariate selection methods include: Gender, Race, Social Economic Status (SES), Kindergarten Math Score, Public School, First-Time Kindergartener, Child's Age at K Entry (Months), Approaches to Learning Rating, Attended Head Start, Received Food Stamps, One-Parent Family and Fine Motor Skills. Most of these covariates coincide with Morgan et al. (2010) summarization of the previously identified strong predictors (i.e., SES (both family and school level), race, gender, learning related behaviors, parents' marital status, reading ability and previous mathematical performances) of the assignment to disability for kindergarten aged children and their mathematical difficulties based on previous research. This illustrates the coincidence between subjective knowledge-based selection and empirical based covariate selection.

Nevertheless, empirically based covariate selection methods drop certain covariates when identified by subjective experts. It is possible to evaluate how these dropped covariates affect causal estimation by comparing the covariate selected by different selection methods from the

change of the point estimates and confidence interval. The IAMB, MMPC, conditional random forests, permutation based XGBoost, Adaptive Lasso and Double Lasso do not largely change the point estimates or narrow the confidence interval, which is a good sign that indicates these selection strategies perform well. Consensus selection achieve the most stable performance across different estimation strategies.

For permutation-based random forest, BART, CTMLE-BART and TMLE-SL's point estimates shift from around -4 to around -2. The confidence interval shifts toward 0. By comparing permutation based random forest with conditional random forest and permutation based XGBoost, which all belong to tree-based selection methods, permutation based random forest including RIRT and P1FSTAMP are not identified by other tree-based approaches. These phenomena are due to the fact that the covariates are marginally independent of the fifth-grade mathematic scores that were explored in Chapter 4.

5.2 Implications of the Study's Results

Feature selection plays an important role in causal estimation. The primary take away from this study is that a researcher should understand the objective of feature selection, the characteristics of the data set, and the difference between causal estimation strategies. Subjective knowledge is essential and always should be used to guide feature selection and causal estimation. Furthermore, causal inference is difficult being that it starts with a large data set with rich covariate sets. By using a combination of empirical and subjective screening, researchers will increase their chance of drawing a more valid causal conclusion. Researchers could follow the suggestions provided by this study and choose relevant covariate selection methods and estimation strategies based on sample size, dimensionality of covariate sets and complexity of assignment/response

surfaces. If there are any disagreements with empirical versus subjective selections, the researcher will rely on subjective guidance.

5.3 Thoughts on Further Research

In this study, the weak versus strong confounder was set for both a large and small sample by using the same coefficients (i.e., 1.5 as a strong relationship and 0.5 as a weak relationship). A more systematic way to set the weak and strong confounders based on sample size and dimensionality of the covariate set is needed.

In this study, only the constant treatment effect has been considered. In fact, under most settings, the treatment effect may be heterogeneous. Different treatment may have a different causal effect on each subject. Estimation of treatment heterogeneity is critical to: (1) identify the most efficient treatment effect; (2) design individual-based treatments for each individual or group; and (3) understand how treatment varies and generalize the treatment effect to the specific target population. The confounding bias and the treatment heterogeneity bias are the two parts that researchers would need to address. There are several simulation and case studies that have been conducted to evaluate the performance of regularization approaches when heterogeneous treatment effect is present (Kern et al., 2016; Wendling et al., 2018). At the same time, to the best of my knowledge, there are just a few researchers, for example, Wu and Holmes (2019) and Chen and Keller (2019) that have developed methods to identify risk factors that result in treatment heterogeneity. One question for further research is how can the risk factor that causes heterogeneity be identified by covariate selections methods? Another important issue to explore is how to further identify "strong" heterogeneity factors that is statistically and empirically important.

Following my extensive simulation study, readers may realize that it is a complicated decision-making process for causal estimation. Different covariate selection methods crossed over with

various estimation approaches will end with dozens of estimation strategies. There are two ways to simplify these procedures: automatic machine learning and machine learning pipeline. Automated machine learning (AutoML) represents a fundamental shift from traditional machine learning estimation. The characteristics of AutoML include automatic feature engineering, automatic model selection and automatic hyper-parameter optimization (Kotthoff et al., 2017). The machine learning pipelines are widely used to help automate machine learning workflow. It provides a step by step procedure to facilitate the efficiency and accuracy of the learning algorithms. In causal inference, it is worth further consideration to combine different feature selection and estimation approaches to develop a causal learning pipeline and automatic the learning procedure. For example, researchers can consider incorporating DWR feature selection library into TMLE super-learning framework to automate the selection and estimation procedure.

Bibliography

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340-1347.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in medicine*, *26*(4), 734-753

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608-650.

Bhattacharya, J., & Vogt, W. B. (2007). Do instrumental variables belong in propensity scores?

Bickel, P. J., Li, B., Tsybakov, A. B., van de Geer, S. A., Yu, B., Valdés, T., ... & van der Vaart, A. (2006). Regularization in statistics. *Test*, *15*(2), 271-344.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Brookhart, M. A., Stürmer, T., Glynn, R. J., Rassen, J., & Schneeweiss, S. (2010). Confounding control in healthcare database research: challenges and potential approaches. *Medical care*, *48*(6 0), S114.

Chen, J., & Keller, B. (2019). Heterogeneous subgroup identification in observational studies. *Journal of Research on Educational Effectiveness*, *12*(3), 578-596.

Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1-4.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, *107*(5), 261-65.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "Bayesian ensemble learning." *Advances in neural information processing systems*. 2007.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266-298.

De Luna, X., Waernbaum, I., & Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, *98*(4), 861-875.

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, *95*(3), 932-945.

Ding, P., VanderWeele, T. J., & Robins, J. M. (2017). Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*, *104*(2), 291-302.

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2017). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *arXiv preprint arXiv:1707.02641*.

Dorie, V. (2020). bartCause: Causal Inference using Bayesian Additive Regression Trees. *R package version 1.0-3*.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, *109*(507), 991-1007.

Ertefaie, A., Asgharian, M., & Stephens, D. A. (2018). Variable selection in causal inference using a simultaneous penalization method. *Journal of Causal Inference*, *6*(1).

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348-1360.

Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology*, *15*(3), 413-419.

Greenland, S. (2008). Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology*, *167*(5), 523-529

Hansen, B. B., Fredrickson, M., Fredrickson, M. M. M., Rcpp, L., & Rcpp, I. (2019). Package 'optmatch'. *Available on https://cran. r-project. org/web/packages/optmatch/optmatch. pdf (last accessed on 10 October 2015)*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Häggström, J. (2018). Data-driven confounder selection via Markov and Bayesian networks. *Biometrics*, *74*(2), 389-398.

Hahn, P. R., Carvalho, C. M., Puelz, D., & He, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, *13*(1), 163-182

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217-240.

Hill, J., Weiss, C., & Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, *46*(3), 477-513.

Hoffmann, K., Pischon, T., Schulz, M., Schulze, M. B., Ray, J., & Boeing, H. (2008). A statistical test for the equality of differently adjusted incidence rate ratios. *American journal of epidemiology*, *167*(5), 517-522.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945-960.

Janzing, D. (2019). Causal regularization. In *Advances in Neural Information Processing Systems* (pp. 12704-12714).

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Machine Learning Proceedings 1994* (pp. 121-129). Morgan Kaufmann.

Jiao, J., Courtade, T. A., Venkat, K., & Weissman, T. (2015). Justification of logarithmic loss via the benefit of side information. *IEEE Transactions on Information Theory*, *61*(10), 5357-5365.

Ju, C., Gruber, S., Lendle, S. D., Franklin, J. M., Wyss, R., Schneeweiss, S., & van der Laan, M. J. (2016). Scalable collaborative targeted learning for large scale and high-dimensional data. *UC Berkeley Division of Biostatistics Working Paper Series. Working Paper*, *352*.

Ju, C., Benkeser, D., & van Der Laan, M. J. (2020). Robust inference on the average treatment effect using the outcome highly adaptive lasso. *Biometrics*, *76*(1), 109-118.

Ju, C., Wyss, R., Franklin, J. M., Schneeweiss, S., Häggström, J., & van der Laan, M. J. (2019). Collaborative-controlled LASSO for constructing propensity score-based estimators in high-dimensional data. *Statistical methods in medical research*, *28*(4), 1044-1063.

Kennedy, E. H., & Balakrishnan, S. (2018). Discussion of "Data-driven confounder selection via Markov and Bayesian networks" by Jenny Häggström. *Biometrics*, *74*(2), 399-402.

Keller, B. (2020) Variable Selection for Causal Effect Estimation: Conditional Random Forest Variable Importance Under Permutation *Journal of educational statistics*.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of research on educational effectiveness*, *9*(1), 103-127.

Khan, M., & Quadri, S. M. K. (2013). Effects of using filter based feature selection on the

performance of machine learners using different datasets. *BVICA M's International Journal of Information Technology*, *5*(2), 597

Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., & Leyton-Brown, K. (2017). Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *The Journal of Machine Learning Research*, *18*(1), 826-830.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1-2), 273-324.

Liu, H., & Motoda, H. (2008). Less is more. *Computational methods of feature selection*, 16-31

Maldonado, G., & Greenland, S. (1993). Simulation study of confounder-selection strategies. *American journal of epidemiology*, *138*(11), 923-936.

Maathuis, M. H., & Colombo, D. (2015). A generalized back-door criterion. *The Annals of Statistics*, *43*(3), 1060-1088.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, *9*(4), 403.

Middleton, J. A., Scott, M. A., Diakow, R., & Hill, J. L. (2016). Bias amplification and bias unmasking. *Political Analysis*, *24*(3), 307-323.

Mickey, R. M., & Greenland, S. (1989). The impact of confounder selection criteria on effect estimation. *American journal of epidemiology*, *129*(1), 125-137.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of special education*, *43*(4), 236-254.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., & Glynn, R. J. (2011). Myers et al. Respond to "understanding bias amplification". *American journal of epidemiology*, *174*(11), 1228-1229.

Neyman, J. (1923). On the application of probability theory to agriculture experiments. Essays on Principles. *Statistical Science, 5(4)* :456-480.

Ning, Y., Peng, S., & Imai, K. (2017). High dimensional propensity score estimation via covariate balancing.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural computation*, *15*(6), 1191-1253.

Parast, L., McCaffrey, D. F., Burgette, L. F., de la Guardia, F. H., Golinelli, D., Miles, J. N., &

Griffin, B. A. (2017). Optimizing variance-bias trade-off in the TWANG package for estimation of propensity scores. *Health Services and Outcomes Research Methodology*, *17*(3-4), 175-197.

Pearl, J. (1988). Probabilistic Inference in Intelligent Systems. San Matteo.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669-688.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2010). An introduction to causal inference. *The international journal of biostatistics*, *6*(2).

Pearl, J. (2011). Invited commentary: understanding bias amplification. *American journal of epidemiology*, *174*(11), 1223-1227.

Persson, E., Häggström, J., Waernbaum, I., & de Luna, X. (2017). Data-driven algorithms for dimension reduction in causal inference. *Computational Statistics & Data Analysis*, *105*, 280-292.

Polley, E. C., & Van Der Laan, M. J. (2010). Super learner in prediction.

Ridgeway, Greg, Dan McCaffrey, Andrew Morral, Beth Ann Griffin, Lane Burgette, Maintainer Lane Burgette, McCaffrey Ridgeway, and Morral Burgette. "Package 'twang'." (2020).

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, *2*(1), 1-26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34-58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, *75*(371), 591-593.

Sauer, B. C., Brookhart, M. A., Roy, J., & VanderWeele, T. (2013). A review of covariate selection for non-experimental comparative effectiveness research. *Pharmacoepidemiology and drug safety*, *22*(11), 1139-1145.

Schapire, Robert E. "Using output codes to boost multiclass learning problems." *ICML*. Vol. 97. 1997.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, *13*(4), 279.

Scheines, R. (1997). An introduction to causal inference.

Schuler, M. S., & Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American journal of epidemiology*, *185*(1), 65-73.

Schisterman, E. F., Cole, S. R., & Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, *20*(4), 488.

Schnitzer, M. E., Lok, J. J., & Gruber, S. (2016). Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. *The international journal of biostatistics*, *12*(1), 97-115.

Scutari, M. (2010). bnlearn: Bayesian network structure learning. *R package*

Sekhon, J. S., & Sekhon, M. J. S. (2020). Package 'Matching'.

Shortreed, S. M., & Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, *73*(4), 1111-1122.

Sobel, M. E. (2005). Discussion: 'the scientific model of causality'. *Sociological Methodology*, *35*(1), 99-133.

Spirtes, P., Glymour, C., & Scheines, R. (1993). Causation, Prediction, and Search, volume 81 of. *Lecture notes in statistics*.

Spirtes, P. (2001). An Anytime Algorithm for Causal Inference. In *AISTATS*

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological methods*, *15*(3), 250.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Tsamardinos, I., Aliferis, C. F., Statnikov, A. R., & Statnikov, E. (2003, May). Algorithms for large scale Markov blanket discovery. In *FLAIRS conference* (Vol. 2, pp. 376-380).

VanderWeele, T. J., & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, *67*(4), 1406-1413.

VanderWeele, T. J. (2019). Principles of confounder selection. *European journal of epidemiology*, *34*(3), 211-219.

Van Der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *The international journal of biostatistics*, *2*(1).

Van der Laan, M. J., & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics*, *6*(1).

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, *6*(1).

Van der Laan, M. J., & Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Vansteelandt, S., Bekaert, M., & Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical methods in medical research*, *21*(1), 7-30.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228-1242.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., & Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in medicine*, *37*(23), 3309-3324.

Wilson, A., & Reich, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics*, *70*(4), 852-861

Witte, J., & Didelez, V. (2019). Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal*, *61*(5), 1270-1289.

Wu, C. H., & Holmes, C. C. (2019). Supervised variable selection in randomised controlled trials prior to exploration of treatment effect heterogeneity: an example from severe malaria. *arXiv preprint arXiv:1901.03531*.

Yu, Kui, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhaolong Ling, and Xindong Wu. "Causality-based Feature Selection: Methods and Evaluations." *arXiv preprint arXiv:1911.07147*(2019).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301-320.

Zou, Hui. "The adaptive lasso and its oracle properties." *Journal of the American statistical association* 101.476 (2006): 1418-1429.

Zhao, Q., Keele, L. J., & Small, D. S. (2019). Comment: Will Competition-Winning Methods for Causal Inference Also Succeed in Practice?. *Statistical Science*, *34*(1), 72-76.

Zheng, W., & van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning* (pp. 459-474). Springer, New York, NY.

Appendix A

Covariates Name	Descriptions	Values
Demographic		
GENDER	Male	0,1
WKWHITE	White	0,1
WKSESL	Socioeconomic Status	[-4.8,2.8]
Academic		
RIRT	Kindergarten Reading Score	[-23.17,139.36]
MIRT	Kindergarten Math Score	[11.9,99.0]
S2KPUPRI	Public School	0,1
P1EXPECT	Parental Expectations	1,2,3,4,5,6
P1FIRKDG	First-Time Kindergartener	0,1
P1AGEENT	Child's Age at K Entry (Months)	[54,79]
ApprchT1	Approaches to Learning Rating	1,2,3,4
P1HSEVER	Attended Head Start	0,1
GHGSCH	Ever Changed Schools	0,1
School Decomposition		
Avg_RIRT	Reading IRT	[27.9,80.0]
Avg_MIRT	Math IRT	[16.1,66.1]
Avg_SES	SES	[-2.2,2.5]
Avg_apprchT1	Approaches to Learning	[1.5,4.0]
S2KMINOR	Percent Minority Students	1,2,3,4,5
Family Information		
P1FSTAMP	Received Food Stamps	0,1,
ONEPARENT	One-Parent Family	0,1
STEPPARENT	Stepparent Family	0,1
P1NUMSIB	Number of Siblings	[1,10]
P1HMAFB	Mother's Age at First Birth	[12,45]
WKCAREPK	Nonparental Pre-K Child Care	0,1
Health		
P1EARLY	Number of Days Premature	[0,112]
P1WEIGHO	Birth Weight (Ounces)	[17,214]
C1FMOTOR	Fine Motor Skills	0,1,2,3,4,5,6,7,8,9
C1GMOTOR	Gross Motor Skills	0,1,2,3,4,5,6,7,8
Parent Rating of Child		
P1HSCALE	Overall Health	1,2,3,4,5
P1SADLON	Sad/Lonely	1,2,3,4
P1IMPULS	Impulsive	1,2,3,4

P1ATTENI	Attentive	1,2,3,4
P1SOLVE	Problem Solving	1,2,3,4
PSPRONOU	Verbal Communication	1,2,3,4
P1DISABL	Child has Disability	0,1
Exposure		
F5SPECS	Special Education Services	0,1
Outcome		
C6R4MSCL	Fifth Grade Math Score	[50.9,170.7]